MULTI-PORT MEMORY DESIGN FOR ADVANCED COMPUTER ARCHITECTURES

by

Yirong Zhao

Bachelor of Science, Shanghai Jiaotong University, P. R. China,

2011

Submitted to the Graduate Faculty of

the Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This thesis was presented

by

Yirong Zhao

It was defended on

July 28, 2013

and approved by

Kartik Mohanram, Ph.D., Associate Professor, Department of Electrical and Computer Engineering

Steven P. Levitan, Ph.D., Professor, Department of Electrical and Computer Engineering.

Alex K. Jones, Ph.D., Associate Professor, Department of Electrical and Computer Engineering.

Helen Li, Ph.D., Assistant Professor, Department of Electrical and Computer Engineering.

Thesis Advisor: Kartik Mohanram, Ph.D., Associate Professor, Department of Electrical and Computer Engineering

MULTI-PORT MEMORY DESIGN FOR ADVANCED COMPUTER ARCHITECTURES

Yirong Zhao, M.S.

University of Pittsburgh, 2013

In this thesis, we describe and evaluate novel memory designs for multi-port on-chip and off-chip use in advanced computer architectures. We focus on combining multi-porting and evaluating the performance over a range of design parameters. Multi-porting is essential for caches and shared-data systems, especially multi-core System-on-chips (SOC). It can significantly increase the memory access throughput. We evaluate FinFET voltage-mode multi-port SRAM cells using different metrics including leakage current, static noise margin and read/write performance. Simulation results show that single-ended multi-port FinFET SRAMs with isolated read ports offer improved read stability and flexibility over classical double-ended structures at the expense of write performance. By increasing the size of the access transistors, we show that the single-ended multi-port structures can achieve equivalent write performance to the classical double-ended multi-port structure for 9% area overhead. Moreover, compared with CMOS SRAM, FinFET SRAM has better stability and standby power. We also describe new methods for the design of FinFET current-mode multi-port SRAM cells. Current-mode SRAMs avoid the full-swing of the bitline, reducing dynamic power and access time. However, that comes at the cost of voltage drop, which compromises stability. The design proposed in this thesis utilizes the feature of Independent Gate (IG) mode FinFET, which can leverage threshold voltage by controlling the back gate voltage, to merge two transistors into one through high-V_t and low-V_t transistors. This design not only reduces the voltage drop, but it also reduces the area in multi-port current-mode SRAM design. For off-chip memory, we propose a novel two-port 1-read, 1-write (1R1W) phase-

change memory (PCM) cell, which significantly reduces the probability of blocking at the bank levels. Different from the traditional PCM cell, the access transistors are at the top and connected to the bitline. We use Verilog-A to model the behavior of $Ge_2Sb_2Te_5$ (GST: the storage component). We evaluate the performance of the two-port cell by transistor sizing and voltage pumping. Simulation results show that pMOS transistor is more practical than nMOS transistor as the access device when both area and power are considered. The estimated area overhead is 1.7×, compared to single-port PCM cell. In brief, the contribution we make in this thesis is that we propose and evaluate three different kinds of multi-port memories that are favorable for advanced computer architectures.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 INTRODUCTION

The goal of this thesis is to describe and evaluate novel memory structures based on multi-porting, new device, technologies used to implement on-chip and off-chip memories, improving their performance through tradeoffs between speed, power, data retention, and stability.

1.2 BACKGROUND, CHALLENGES AND MOTIVATION

The memory system of a computer includes many levels of memories. They can be divided into two groups, on-chip memories and off-chip memories. On-chip memories include the instruction cache and the data cache, which are mostly realized in SRAM. Off-chip memories are the main memory, which are mostly realized with DRAM, and sometimes L3 caches due to their larger capacity and size. However, traditional CMOS SRAM and DRAM have their limitations as technology scales below 22nm and more requirements are placed on design for the next generation [5].

1.2.1 Challenges

In next generation, there are three major concerns for memories, which are low cost, transistor scaling and non-volatility. The cost of a chip is related to many factors: the power consumption, the area, the yield in fabrication process and the speed to read and write

data. In traditional CMOS SRAM, there should be a trade-off between the first two factors. Scaling down of transistors can reduce the cost because it increases transistor density, reduces power consumption and improves performance by reducing the gate delay. However, it also brings a severe problem of short channel effects (SCE). Punch through between drain and source and drain-induced barrier lowering (DIBL), surface scattering and velocity saturation are all examples of SCE. These problems reduce the on-current, increase the sub-threshold current and decrease the threshold voltage, which degrade the cell stability to create errors and increase leakage power [6]. In addition, the yield can be reduced because the effect of process variation is more prominent since variation becomes a larger percentage of the full length or width of the device [7]. For the main memory, volatility is the most critical problem for traditional DRAM because it only relies on the capacitor in its cell structure to store the value. The system should refresh every several microseconds to maintain data integrity. Besides the storage capacitor, the sub-threshold charge leakage should also be mitigated for the access transistor. As a result, new technology and new memory structures are being actively investigated to meet the requirements for the next generation of memories.

1.2.2 State-of-the-art

There are several modern technologies that are being investigated. For example, spin-transfer torque RAM (STT-RAM) and phase change memory (PCM) are novel non-volatile off-chip memories. They both use new materials as the storage component in the cell. For on-chip memories, FinFET SRAM and SRAM with Si tunnel transistor [8] are designed to replace the traditional CMOS transistor used in the SRAM, with enhanced cell stability and lower power consumption. Furthermore, several SRAMs with different cell structures have been also designed to meet some of the requirements of the next generation [5]. Some SRAMs are developed based on the conventional 4T loadless SRAM cell, wherein the two upper pMOS transistors are directly tied to the bitlines instead of the supply voltage as the access transistors. Inheriting the advantages of reduced area and power consumption of 4T, the proposed 5T [9] and 7T [10] SRAM cells both have enhanced read stability. Another novel structure is the Schmitt trigger SRAM [11]. This SRAM utilizes the principle

of Schmitt Trigger, which has different threshold voltages between the falling edge and the rising edge of input. Therefore, the disturbance of one internal node in that SRAM should be larger than that in the classical 6T SRAM to make the other internal node to toggle its value. That makes the SRAM achieve 1.56× higher read static noise margin. Some of those novel structures are already implemented with IG mode FinFET in the literature [12, 11] to reduce the area and improve the stability.

1.2.3 Multi-porting: Background and motivation

These works we have introduced are all single-port memories. But with the emerging popularity of the multi-core processors for commodity computing, since every core has to communicate with the memory and meanwhile the correctness of data store and read of the memory should be ensured, multi-porting is considered an efficient way to reduce the wait time for different cores for the same memory so as to reduce the throughput of the access time of each core. A quad-port shared memory based on FPGAs was proposed [13], which increases the speed of write and read by almost 66.7%, compared to a single-port memory. Multi-port register files were implemented in the AMD K7 processor [14] and the Itanium Microprocessor [15]. However, the technology of multi-porting and those new technologies and cell structures have not been investigated extensively in the literature. Those modern techniques such as PCM and STT-RAM are known for their data retention, endurance, stability and scalability or cell density, but they could not bring large improvement on their data access time. For example, to write to a PCM cell, we must change the supply voltage and the driver voltage to a suitable value, and then maintain those voltages to heat the GST (the storage component) until it changes its status. Usually it takes 100ns to 1000ns [16] for write operation and 10ns to 100ns for read operation. It is smaller in read operation because the bitline and wordline does not need to be pumped to a higher voltage. This time is much greater than the traditional 6T SRAM and DRAM which are both around 10ns. Even for the SRAM, the change of structure, such as Schmitt Trigger SRAM, substantially contributes to the stability of SRAM, does not do favor to the read/write acceleration. To address these issues, multi-porting is a potential solution. Multi-porting enables several

read/write requests served at different memory addresses in different memory cells almost simultaneously, which is unlike that only one request can be served at one address in one cell of the single-port memories. Therefore, read/write speed or throughput will have a considerable increase for those modern cell structures. For PCM, it is more convenient to divide write from read because their bitline voltages and the driver voltages are totally different. The most common multi-porting is two-porting, i.e. one write port and one read port in a cell.

1.3 CONTRIBUTIONS

This thesis makes the following contributions. First, we perform the study of multi-port FinFET SRAM cells. To the best of our knowledge, this is the first work evaluating read/write acceleration through multi-porting in FinFET SRAMs. Since multi-port SRAMs can be fundamentally classified into two types, double-ended multi-port SRAMs and single-ended multi-port SRAMs, we include one double-ended structure and three single-ended structures with two isolated read ports in our thesis. We evaluate those cells and the classical 6T SRAM as a baseline using different metrics, such as leakage current, static noise margin, read/write time with the predictive technology model (PTM), and compare them with each other. Based on simulation results, we can conclude that single-ended multi-port FinFET SRAM with isolated read ports has the advantage over the double-ended multi-port FinFET SRAM in the read operation, such as robust read access and high flexibility in read port configuration. However, its drawback is that its write performance is worse than the double-ended structure, due to its weakness in breaking the feedback loop in write operations. Finally, we show that the write performance of single-ended structures with one read port can be improved to the same level of the classical double-ended structure for 9% area overhead by sizing the access transistors appropriately.

Second, we propose novel structures of multi-port current-mode SRAM using the technique of FinFET to deal with the voltage drop issue in the traditional multi-port current-mode SRAM without any area overhead. Due to the ability of controlling its back gate

voltage, which can create feedback, and merging transistors, FinFET can reduce cell size and improve performance in the multi-port SRAM design. Our results show that the voltage drop can be reduced by around 20% of the full CMOS voltage supply when the transistors are sized normally i.e., W=40nm for nMOS and W=80nm for pMOS. Further, by merging transistors using IG mode FinFET in the two-port current-mode structure, area can be substantially reduced.

Finally we propose a new two-port phase change memory cell as an off-chip main memory cell substituting traditional DRAM cell structure, which significantly reduces the probability of blocking at the bank and architecture levels. The blocking is caused by the asymmetry in read/write access latency, which leads to the low throughput performance, and impedes non-volatile memory to be integrated into high-performance computing systems. We design a model of the GST and we come up with a series of suitable bitline and wordline voltages in its write and read operations. The innovation is that the access transistors are on the top instead of on the bottom to better construct a two-port cell. Result shows that pMOS transistors are more practical than nMOS transistors as the access device when both area and power are considered. The innovation can reduce the expected read delay by 12-40× over conventional single-port PCM for 1.7× overhead. This thesis is organized as follows.

1.4 THESIS ORGANIZATION

In Chapter 2, we will introduce the background of FinFET SRAM design and present our work on multi-port voltage-mode FinFET SRAM design. We will present in detail the metrics we use to evaluate and compare the performance of single-ended and double-ended multi-port SRAM cells. In Chapter 3, we will introduce the techniques of IG mode FinFET and its use in the proposed multi-port current-mode SRAM cell to solve the problem of voltage drop. In chapter 4, we will propose and evaluate the 1R/1W two-port PCM cell. Chapter 5 is the conclusion.

2.0 FINFET MULTI-PORT VOLTAGE-MODE SRAM EVALUATION

In this chapter, we are going to focus on the evaluation of two kinds of multi-port FinFET voltage-mode SRAMs: single-ended and double-ended using different metrics. We will analyze the result of the read and write acceleration by multi-porting compared with the classical 6T SRAM. We will also compare the result of those FinFET SRAMs with corresponding CMOS SRAMs.

2.1 INTRODUCTION OF FINFET

FinFET, a double-gate device has emerged as an alternative of traditional CMOS. Different from CMOS, it has two gates to control the channel, front gate and back gate. They are electrically coupled to better control the short channel effect, which substantially lower the DIBL and increase the subthreshold slope, i.e. the on/off current ratio. Therefore, it is a better choice for its application in low power SRAM circuits. In addition, it has a body called 'fin' which is perpendicular to the substrate and is generally enclosed by oxide. It is thin and lightly doped, enabling FinFET to further suppress the SCEs and control the process variation of the threshold voltage, which is a serious problem in bulk Si CMOS. Basically, high on/off current ratio is the most advantage of FinFET over CMOS, due to the better control over SCE, which provides opportunities to reduce standby power [17]. With this advantage, in FinFET SRAM and other FinFET circuit design, people usually use back gate control method to further improve the performance. It is a unique method for FinFET because CMOS only has one gate. In FinFET device, the front gate and back gate can be tied or untied. They correspond to two working mode of FinFET, Shorted Gate (SG) mode

and Independent gate (IG) mode. SG mode enables a much larger on current of FinFET than IG mode. However, IG mode has more flexibility. In IG mode, we are able to control the back gate voltage. Higher back gate voltage produces higher threshold voltage and lower back gate voltage produces lower threshold voltage. The back gate voltage can be as low as -0.26V. If we want the circuit have high stability and low leakage current, we can set a high threshold voltage. And if we want the circuit to have short access time, we cab set a low threshold voltage. These properties cannot be found in CMOS. Many current creative works in FinFET SRAM design involves back gate control. In [13], the upper front FinFET transistor works in IG mode and its back gate voltage is set to zero to provide a high threshold voltage so that the circuit has a large retention and read static noise margin.In [18], a footer is also designed using IG mode FinFET transistor to control the standby power of the SRAM cell in read, write and retention state. In [12] and [19], IG mode FinFET is also implemented in novel 4T and 7T SRAM structure. In Schmitt Trigger SRAM, FinFET in IG mode also has its application [11]. The saving of two transistors is the benefit of this method. Thus FinFET is an ideal alternative to planar CMOS in SRAM design.

2.2 INTRODUCTION OF MULTIPORT VOLTAGE-MODE SRAM

There are generally two types of cell structure in voltage-mode multi-port SRAMs: the double-ended structure and the single-ended structure. Double-ended structures have two bit lines on each port and single-ended structures have only one. Figure 1 is the most common double-ended two-port SRAM. Figure 2 is three single-ended SRAMs with two isolated read ports. The double-ended structure can easily break the feedback loop in the write operation using its two bit lines. Meanwhile, the single-ended structure has a short wire delay and a small cell area, due to its single-ended bit line and word line [1]. As SRAM cells are scaled down, single-ended multi-port SRAMs and register files with isolated read ports become more attractive. For instance, a single-ended multi-port register file was built in the Itanium Microprocessor [15], and a single-ended 34 bit $\times$ 64 bit 10R/6W register file was proposed [20]. However, to avoid the longer write latency in the single-ended structure [1],

the double-ended structure is preferred in multi-port SRAM design to provide faster access, e.g., the register file in the AMD K7 processor [14] and the 8T double-ended two-port SRAM cell [21]. In this paper, single-ended cells with isolated read ports are referred to as single-ended structures. Note that the double-ended structure with isolated read ports in not considered due to its large area overhead.

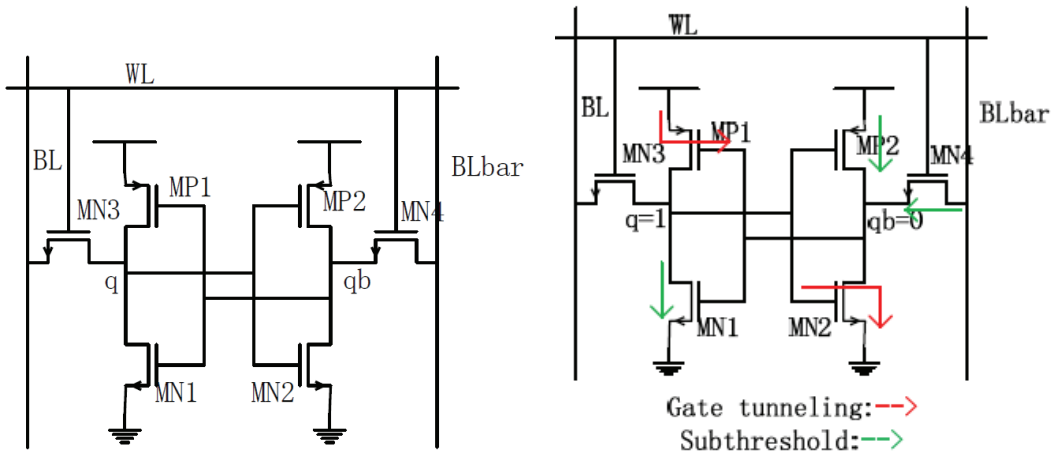
To accelerate read operations in a double-ended structure SRAM cell, we can duplicate access transistors to compose more ports. Thus, we can read the value in one cell from multiple ports simultaneously when they are enabled. The total access time of two simultaneous reads in the same row but not in the same cell can be reduced largely when two word lines are enabled at the same time. However, the challenge is that the total access time of two simultaneous reads in the double-ended structure is longer than one single read in the single-port cell. This is because adding one read port brings another bit line to the cell, making it more difficult to pull down the pre-charged bit lines, due to the larger capacitance.As the number of ports increases, the read access time for simultaneous read grows. Single-ended structures, however, do not suffer from such a problem. Due to isolated read ports, the cell provides gate voltage to read port transistors while read port transistors do not affect the cell. The bit line is only driven by the read port transistors instead of the cell. Since read ports are isolated with each other, they only need to drive their own bit line, providing equivalent driving ability for simultaneous multiple reads to one single read. As a result, increasing ports does not increase the simultaneous read time. Furthermore, the destructive read problem gets worse in simultaneous read operations in double-ended structures. Adding access transistors in double-ended structures makes the internal node affected seriously by the supply voltage of every bit line. Therefore, the single-ended structure is better than the double-ended structure in terms of read acceleration.

For write acceleration in the double-ended structure, duplicating access transistors is the preferred approach. However, the time of one write slightly increases in multi-port cells because the extra access transistor adds extra capacitance to the internal node. The simultaneous write can occur in the same row, but not in the same cell, because it is impossible to write multiple values into one cell in one cycle. Thus, the total write time for two simultaneous writes in the same row but different cells is close to the time needed for one

single write in the single-port structure, while the total write time for two writes in the same cell still remains the same as the total time for two writes in the single-port structure. For the single-ended structure, the critical issue in write acceleration is the need of write-assist structures to reduce the write access time, because the feedback loop in the single-ended structure is strong.

A common challenge in both read and write multi-porting is the half-select problem: for a cell in retention, the word line voltage may be high, when other cells in the same row are being read or written to, which enables the bit line supply voltage to affect the internal node [22]. As the number of ports increases, the cell is more vulnerable to the half-selected problem. To address the issue, a divided word line configuration was introduced to completely eliminate the condition of disturbed access in the unselected cell using hierarchical blocks of word line selection logics [23], to mitigate power and area overhead.

2.3 CLASSICAL 6T FINFET SRAM CELL



(a) (b)

Figure 1: (a) 6T classical SRAM cell [1]. (b) leakage current model for 6T SRAM cell [1].

To begin with the evaluation of above FinFET multi-port SRAM cells, we evaluate the classical single-port 6T FinFET SRAM cell as a baseline. For the FinFET technology, there are two popular models: UFDG model [24] and PTM [25]. We use the double-gate PTM

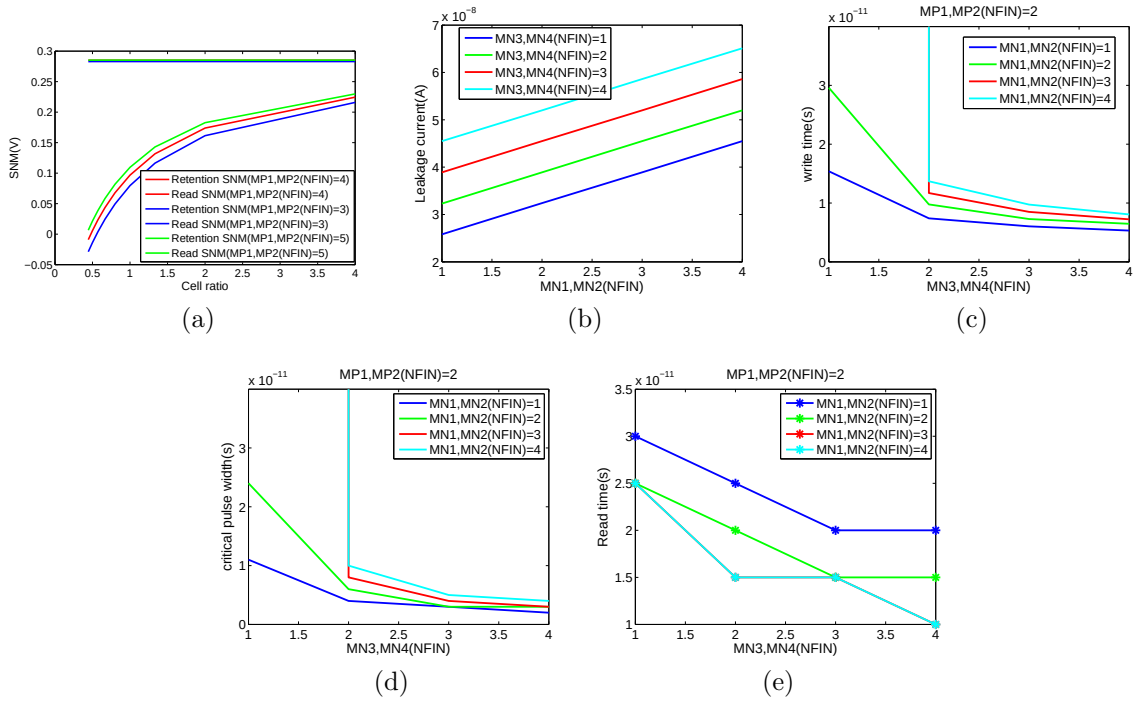


Figure 2: (a) Retention and read SNM of the 6T SRAM cell. (b) Leakage current of the 6T SRAM cell.(c) Write time of the 6T SRAM cell. (d) Critical pulse width of the 6T SRAM cell. (e) Read time of the 6T SRAM cell. (The red curve and the blue curve fully overlap).

22nm technology model and HSPICE as the platform in our FinFET SRAM simulations. In our simulations, based on the equation $W_{eff} = (2 \times h_{fin} + W_{fin}) \times N_{fin}$, where the fin height h_{fin} and fin width W_{fin} is fixed, we can change the number of fins N_{fin} to get a reasonable effective channel width W.

Figure 1(a) illustrates the classical FinFET SRAM cell. Generally, cell stability, power, and read/write performance are common metrics in evaluating an SRAM cell. Cell stability is evaluated in terms of static noise margin (SNM) [1], measuring the DC disturbance in the unselected condition and the read condition. It is obtained by measuring the two squares in butterfly curves [1]. There are mainly two types of SNM that should be considered in SRAM designs: retention SNM and read SNM. The SNM in half selected condition (half-selected SNM) is the same as read SNM in classical FinFET SRAM. The static power, in terms of leakage current, is mainly sub-threshold current and gate tunneling current, as shown in Figure 1(b). In FinFET, the gate tunneling current is negligible, compared to the sub-threshold one. Read performance is evaluated in terms of read time, which is the time between the rise of word line and the time the sense amplifier can detect the voltage difference of the two bit lines [26]. We evaluate the write performance using two metrics: write time and critical pulse width. Write time is the time between half the rise of word line and half the rise of node 'qb' when writing a '0' [18]. Critical pulse width [27] is the smallest word line pulse width to successfully write a '0' or a '1' to the cell.

Cell stability: Figure 2(a) shows the result of retention SNM and read SNM of the classical FinFET SRAM with different cell ratios, which is the ratio of the size of the pull-down transistor to the size of the access transistor [1]. As the cell ratio grows, the read SNM also increases. Figure 2(a) also shows the pull-up transistors have limited impact on read SNM. For retention SNM, it is almost independent to the cell ratio.

Power: Figure 2(b) shows the sum of two types of leakage current in this FinFET SRAM cell, in which the sub-threshold is dominant. The sub-threshold current is proportional to its conductance, which is proportional to the transistor width. Thus, the leakage current is proportional to the width of the transistors.

Read time: In the simulation setup, the bit line capacitance is set based on the estimation of a standard 128×128 SRAM array. Figure 2(c) shows the read time when we change the

size of access transistors. We observe that read time decreases when we increase the width of pull-down transistors (MN1 and MN2), and it increases when we decrease the width of access transistors.

Write performance: As shown in Figure 2(c) and Figure 2(d), when we increase both the widths of pull-down transistors, the write time and the critical write pulse becomes wider, which is contrary to the case of read time. However, when we increase the size of access transistors, both read and write time decreases. When the cell ratio is too small, such as $N_{fin} \geq 3$ for MN1/2 and $N_{fin} \leq 2$ for MN3/4, the values of write time and the critical pulse width are almost infinite, which means the write operation fails. Note that when evaluating the read/write performance, we just change the size of pull-down transistors, instead of changing cell ratio, which is not directly related to the read/write performance. Changing the size of pull-up transistors (MP1 and MP2) has similar impact to sizing the pull-down transistors and is not reported here.

2.4 MULTI-PORT FINFET SRAM CELLS

In this section, we present a comparison of the double-ended and single-ended FinFET multi-port SRAM structures to the classical 6T SRAM.

2.4.1 Double-ended multi-port FinFET SRAM

In a double-ended multi-port SRAM, two extra access transistors are added to the classical FinFET SRAM cell to compose an extra port in the double-ended multi-port FinFET SRAM cell. In this sub-section, we present simulation results of this structure comparing this structure to the 6T cell.

Cell stability: Figure 4(a) shows the result of the retention SNM, the one-read SNM and the two-read SNM when the width of the pull-up transistors is fixed. One-read SNM of a cell is the SNM when there is only one read access served in the column. It may cause the half selected condition if the read is served in another cell. Two-read SNM of a cell is the SNM

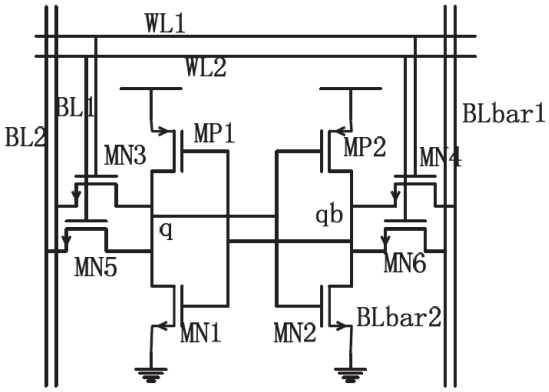


Figure 3: Double-ended multi-port SRAM cell with two sets of access transistors.

when there are a read access is served in this cell while there is another read/write access is served in another port of the same cell or in another cell of the same row. In this case, if there is a read and a write served in the same row, it is a severer half selected condition, because the cell is disturbed simultaneously by the read word line and the write word line. In Figure 4(a), the two-read SNM is much worse than the one-read SNM. Therefore, when we continuously increase the number of access transistors without any additional structures to avoid the half-selected condition, the cell becomes fragile.

Leakage current: Figure 4(b) shows that increasing the width of pull-down transistors or access transistors leads to the growth of the leakage current. Also, the extra two access transistors in the 8T cell compared the 6T cell contribute in higher leakage current.

Write performance and read time: In Figure 4(c) and (d), write time and CPW is slightly larger than that of the 6T FinFET SRAM, because the two extra access transistors brings extra capacitance to the cell, which increases the charging time. The read time of one read in either of the ports is not included in the graph, because it is the same to the read time in the classical 6T FinFET SRAM. However, we include the results of the read time in simultaneous reads scenario in Figure 4(e), compared with the read time in the 6T FinFET SRAM, which shows that the read time gets worse in the double-ended multi-port FinFET SRAM when we simultaneously read in multiple ports.

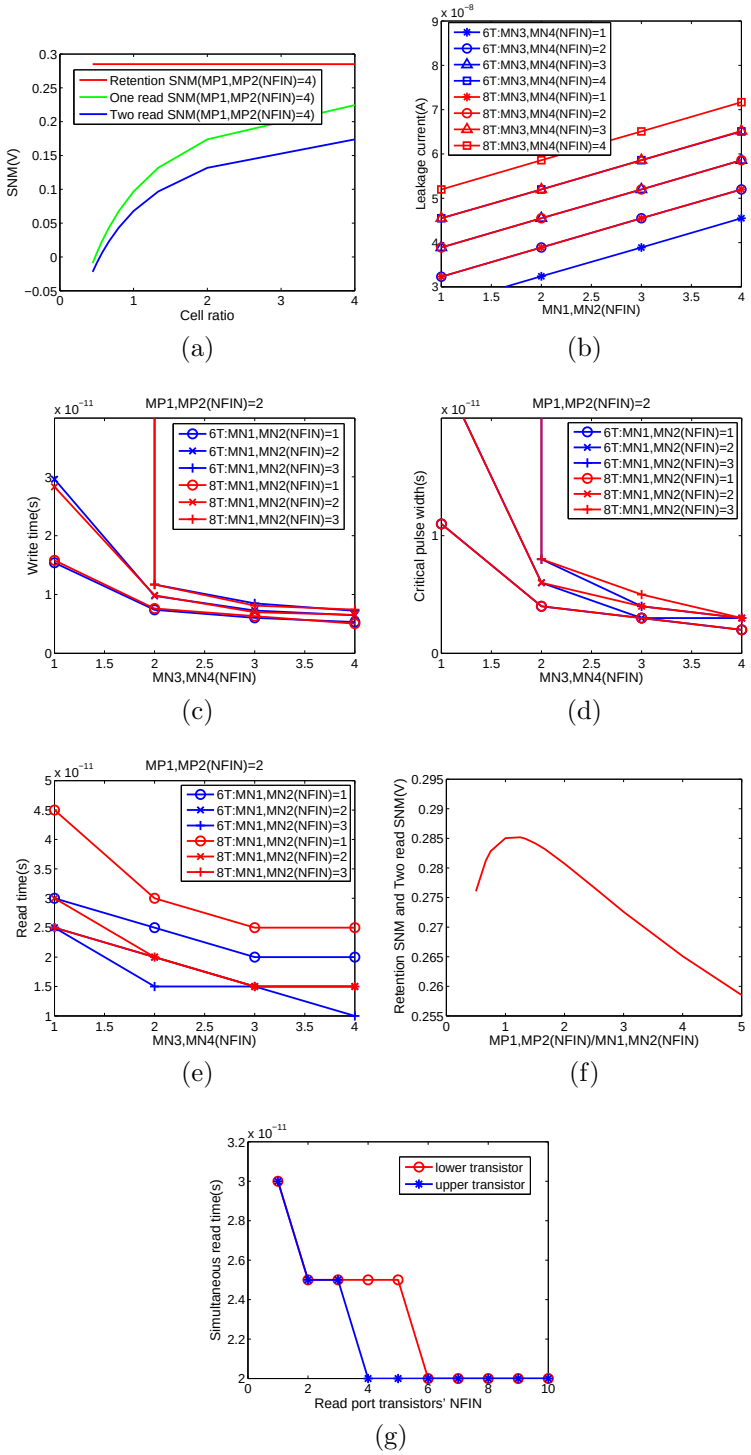


Figure 4: (a) SNM, (b) leakage current, (c) write time, (d) critical pulse width, and (e) simultaneous read time of the double-ended FinFET SRAM cell with one and two sets of access transistors; (f) retention and two read SNM and (g) simultaneous read time of three single-ended structures;

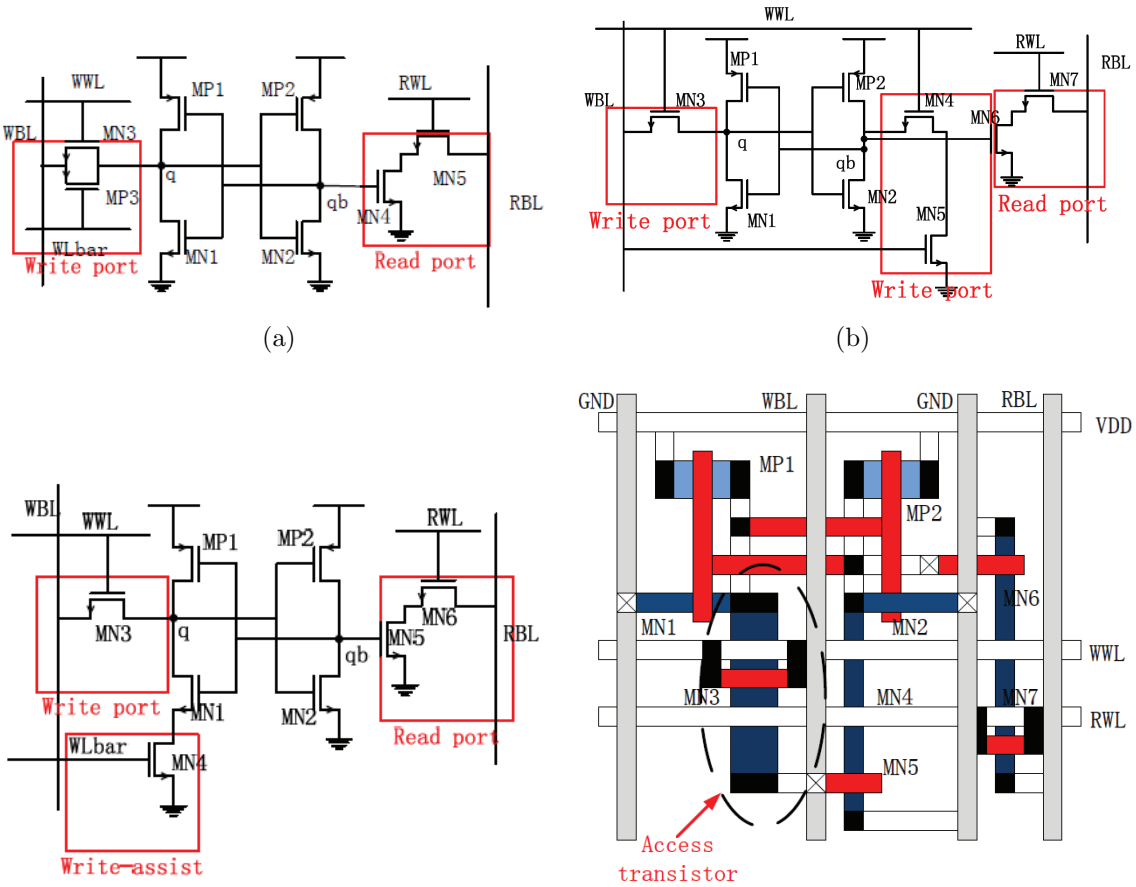


Figure 5: (a) Single-ended multi-port FinFET SRAM cell structure 1 [1]. (b) Single-ended multi-port SRAM cell structure 2 [1, 2]. (c) Single-ended multi-port SRAM cell structure 3 [1]. (d) Layout of structure 2 based on [3].

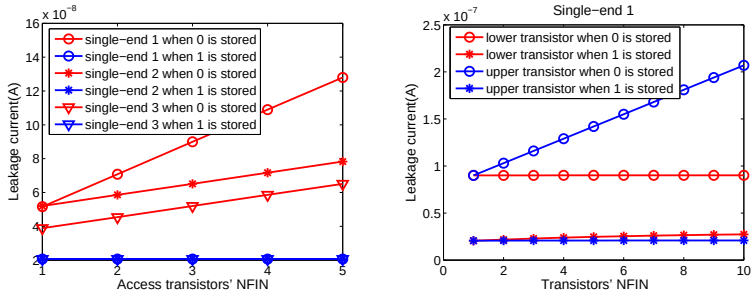
2.4.2 Single-ended multi-port FinFET SRAM

In this sub-section, we introduce three different single-ended multi-port SRAM structures, as shown in Figure 5. Write-assist techniques are implemented in all these three structures to enhance the write performance. There are four combinations for at most two write ports and two read ports: 1R/1W, 1R/2W, 2R/1W, and 2R/2W. We compare the simulation results of 2R/1W and 2R/2W structures, since the performance of structures with one or two read ports is almost the same, except for leakage current, which linearly increases as read ports increase in our simulation.

In Figure 5(a), a transmission gate is used to replace the classical nMOS access transistor, which can write a strong '0' and '1' to the cell. In Figure 5(b), we use MN4 and MN5, two nMOS transistors, to create a pseudo-double-ended structure when writing a '1'. In Figure 5(c), an extra write assist transistor MN4 is added as a switch, which turns off the feedback loop when writing '1'. The extra transistor is shared through a whole row, reducing the area [2].

Cell stability: Since read ports are isolated, the retention SNM and the two-read SNM are the same, regardless of the change in N_{fin} of the other transistors other than the read port transistors. The only parameter that influences the retention SNM and the two-read SNM is the ratio of the size of pull-up transistors to the pull-down transistors. And for these SNMs, three structures have the same performance because they have the identical internal structure. Figure 4(f) shows that the SNM has a peak value, which locates at the ratio range between 1 and 1.5. This is the point that the resistance of the pull-down transistor and the pull-up transistor is almost the same. Half-selected condition can happen in single-ended structures when the write word line is enabled in a whole row, affecting unselected cells. When the voltage of 'qb' is changing from '0' to '1', node 'q' cannot be pulled down to ground due to the impact of the high voltage of the bit line. However, this condition rarely happens, because of no bit line on the side of 'qb' causing the DC disturbance. Therefore, we do not include the half-selected SNM of single-ended structures in this paper.

Read time: Based on simulation results, the only parameter that significantly affect the simultaneous read time is the size of read port transistors. However, the lower transistor



(a) (b)

Figure 6: (a) Leakage current versus the access transistors of the three single-ended structures. (b) Leakage current versus the read port transistors of single-ended structure 1.

and the upper transistor in the read port also have minor effects as reported in Figure 4(g).

Leakage current: It is measured in the scenarios when '0' or '1' is stored, since the cell is asymmetric. We only evaluate the result of changing the size of read port transistors and access transistors in 2R/1W structures, because increasing the write port transistors linearly increases the leakage current. In Figure 6(a) and Figure 6(b), the result shows that when '0' is stored, increasing either the size of the upper read port transistor or the access transistors causes high leakage current. The impact of two internal inverters on leakage current in single-ended structures is similar to that of the double-ended structure. When we add a write port, the leakage current increases by a certain amount when '0' is stored. In the case when '1' is stored, extra write ports do not impact the leakage current, since write port transistors, which have no voltage difference between their two terminals, do not contribute to leakage current.

Critical pulse width (CPW) and write time: They are also evaluated in the scenarios of writing '0' or '1'. The feedback inverter and the forward inverter have different influences in the write performance. For each inverter, we set the size of pMOS transistor to twice the size of the nMOS transistor. In Figure 7(a) and (e), feedback transistors seriously affect the write time and the CPW of structure 2 and 3 when writing '0'. They rapidly increase to infinity as the size of inverter increase. That is because when the feedback inverter is too strong, the access transistor cannot pass '0' to the internal node. Comparatively, the effect when writing '1' is not obvious for structure 2 and 3, because the feedback inverter is off in

structure 3 and pseudo double-ended structure is enabled in structure 2 when writing '1'. In structure 1, the transmission gate can pass either strong '0' or strong '1' at the cost of leakage current. Meanwhile, for the forward inverter, when we change the size of transistors, the change in write performance is less significant than that of the feedback inverter, because there is no conflicts between the effect of the access transistors and the forward inverter. However, as shown in Figure 7(b) and (f), the effect of changing the size of access transistors is opposite to that of feedback inverters on CPW and write time. Finally, in Figure 7(c) and (g), increasing the size of the lower transistor in the read port increases the write time and the CPW, because the gate capacitance is increased. Compared to forward inverters, it has a larger impact on write performance. In addition, if we add an extra write port, the write performance is different between writing '0' and writing '1'. It is better when writing '1' and worse when writing '0'. Figure 7(d) and (h) shows the write time and the CPW of the first two single-ended structures.

2.4.3 Multi-port vs. single-port FinFET SRAMs

In this sub-section, based on comprehensive comparisons, we show that the single-ended multi-port FinFET SRAM provides the best tradeoffs between read acceleration and the read cell stability. Further, for less than 10% area overhead in sizing the access transistors, it can deliver equivalent write time to all other cells.

Leakage current: As double-ended and single-ended multi-port structures have more transistors than the 6T FinFET SRAM cell, they generate more leakage current with the same transistor size. Single-ended structure is a little worse than the double-ended one. However, there is a method for single-ended structure to reduce the leakage current, while maintaining the read time and write time at a reasonable level. That is to decrease the size of feedback inverters to its minimum value. In contrast, for the double-ended FinFET SRAM structure or the classical 6T FinFET SRAM, this approach is not feasible, because the read time and the write time change in opposite directions when we change the internal inverters size, and reducing the size of access transistors increases both the read time and the write time.

Read time: The read time of FinFET SRAM cells is in the order of 10^{-11} s. Single-ended

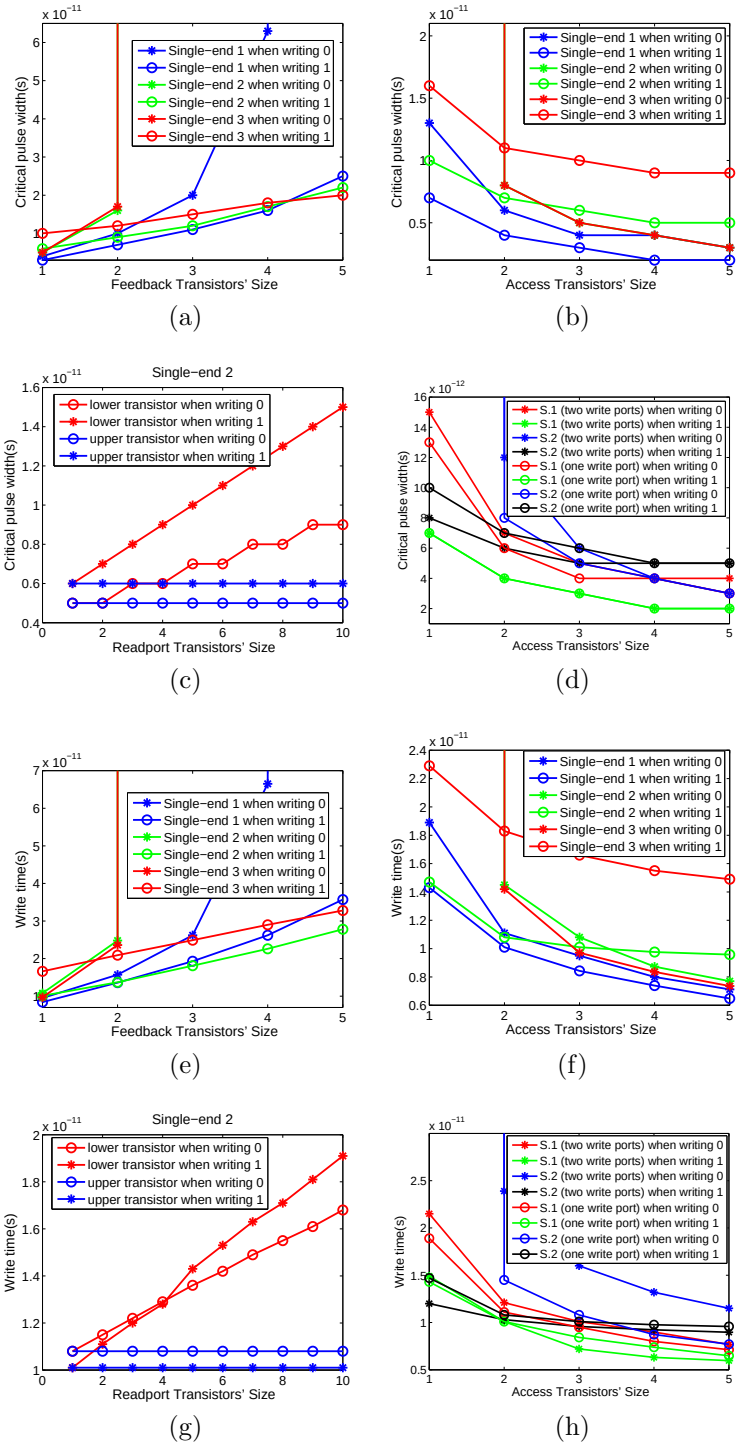


Figure 7: (a) CPW of single-ended structures for feedback inverter sizing; (b) CPW of single-ended structures for access transistor sizing; (c) CPW of structure 2 for read port transistor sizing; (d) CPW of structures 1 and 2 (1–2 write ports) for access transistor sizing; (e) Write time of single-ended structures for feedback inverter sizing; (f) Write time of single-ended structures for access transistor sizing; (g) Write time of structure 2 for read port transistor sizing; (h) Write time of structures 1 and 2 (1–2 write ports) for access transistor sizing.

multi-port FinFET SRAM is the best candidate in multi-port FinFET SRAM structures for getting the similar read time of classical 6T FinFET SRAM cell. There are two advantages of single-ended structures regarding to read operations. The first one is that, regardless of the number of read ports added, the simultaneous read time is identical to that of the single-ended structure with only one read port. The other advantage is that we only need to select proper sizes of transistors for read ports to ensure the reasonable read time in single-ended structure while maintaining leakage current and cell stability. In double-ended structures, the read time is related to internal transistors and access transistors. In order to make the simultaneous read time close to the read time in classical 6T FinFET SRAM cell in the double-ended structure, we should increase the size of access transistor, which seriously increases power and decreases cell stability.

Cell stability: For retention SNM, the result of the multi-port structures is similar to that of 6T FinFET SRAM. The difference is in read SNM. In the single-ended multi-port structures, we can easily achieve the same read SNM as retention SNM, because of the isolated read ports. In the double-ended structure, the simultaneous read SNM is worse, compared to the single-ended structure, due to the isolated read port in the single-ended structure. Meanwhile, in the single-ended structures, the half selected condition decreases its cell stability and make it worse than that of the double-ended structure, because there is almost no DC disturbance on 'qb' node while in the double-ended structure. The half selected SNM is much worse as the number of ports increases in single-ended structures.

Write performance: Both CPW and write time is in the order of 10^{-11} to 10^{-12} s in all the structures simulated with the FinFET models. We observe that in Figure 4(c) and (d), increasing the ports of write and read does not have a significant impact on the write time in both double-ended and single-ended multi-port structures. And we should set the size of access transistors in single-ended structures to 2 to 3 times that of double-ended structures, or make the feedback inverter of single-ended structures even smaller, in order to achieve the same write performance of the double-ended structure. For example, consider the layout of structure 2 with one isolated read based on the classical 6T cell [3] in Figure 5(d), where sizing the access transistors three times adds 9% area overhead.

2.4.4 Comparison with corresponding CMOS SRAM cell

Last but not least, we also did the simulation of the classical 6T SRAM, the double-ended structure and one of the single-ended structure in planar CMOS to obtain the visual result of the improvement of performance that FinFET provides. To make the result more comparable, the model we use is the PTM 22nm high power CMOS model compared with the 20nm length PTM double-gate model used in FinFET multi-port SRAM design. The voltage supply for CMOS model is 0.9V, which is equal to the supply voltage in FinFET multi-port SRAM. The channel width we choose is 66nm, because according to the equation W=(2*Wfin+Tfin)*Nfin and the model data, the channel width of FinFET with a single 'fin' is 70nm.

Table 1: Comparison of the performance of single-port and multi-port SRAMs by using FinFET and CMOS

		CMOS	FinFET
classical 6T SRAM	Retention SNM	0.244V	0.283V
	Read SNM	0.072V	0.097V
	leakage current	6.76×10^{-8}A	2.58×10^{-8}A
double-ended multi-port SRAM	two read SNM	0.010V	0.068V
single-ended multi-port SRAM 1	leakage current('0' stored)	$2.56 \times 10^{-7}A$	9×10^{-8}A
	leakage current('1' stored)	5.49×10^{-8}A	2.06×10^{-8}A
single-ended multi-port SRAM 2	leakage current('0' stored)	1.52×10^{-7}A	6.51×10^{-8}A
	leakage current('1' stored)	5.49×10^{-8}A	2.06×10^{-8}A

Result shows that the read and write performance are not improved a lot, because the total parasitic capacitance of a FinFET transistor is of the same order of magnitude. There-

fore, we do not put it in table 1. However, as table 1 shows, the cell stability (represented by retention and read static noise margin) and the leakage current are improved. That result proves that FinFET can to some degree suppress short channel effect and increase the subthreshold slope.

2.5 SUMMARY

Our evaluations of read and write acceleration in different multi-port FinFET SRAM structures illustrate the impact on read/write performance, leakage current, and cell stability. Based on simulation results with the PTM FinFET model, single-ended multi-port FinFET SRAM with isolated read ports is a good choice for multi-port design, since for similar leakage current, write time, and 9% area overhead, it performs better in read operation, offers higher flexibility in the configuration of read acceleration, and provides better cell stability than double-ended multi-port FinFET SRAM. In addition, we also can conclude that FinFET is a good alternative to CMOS in designing SRAM due to its better stability and lower standby power consumption.

3.0 FINFET MULTI-PORT CURRENT-MODE SRAM

Current-mode SRAMs are advantageous over conventional voltage-mode SRAMs since they avoid the full-swing of the bitlines, improving performance and reducing power. However, this comes at the cost of bitline voltage drop, which compromises cell stability. We propose two methods implemented by IG mode FinFET to solve this problem without adding area overhead. One method is to create a feedback between the bitlines and the access transistors with a structure of NAND function using high-V_t and low-V_t FinFET transistors. The other method is to substitute the SG mode access transistors with IG mode transistors and tie the back gate to the ground. We then evaluate the performance of them and use one of them in the multi-port current-mode SRAM design.

3.1 INTRODUCTION OF CURRENT-MODE SRAM

Using only FinFET in multi-port voltage-mode SRAM design is not enough to reduce the read and write access time of each port within the cell because it cannot reduce the total parasitic capacitance of the cell. The current-mode SRAM rather than traditional voltage-mode SRAM can solve this problem [28]. The most critical factor that influences the access time is the time for the bitline value to be charged and discharged. In writing, one of the bitlines should be charged to '1' and in reading, both of the bitlines should be charged to '1' and then one of the bitlines should be discharged. Fully charging and discharging the large bitline load requires a good amount of time. Using a large driver helps to some extent, but the problem is shifted to the driver. Therefore, current-mode SRAM, which does not need full swing of the bitline, can provide an alternative to achieve these objectives.

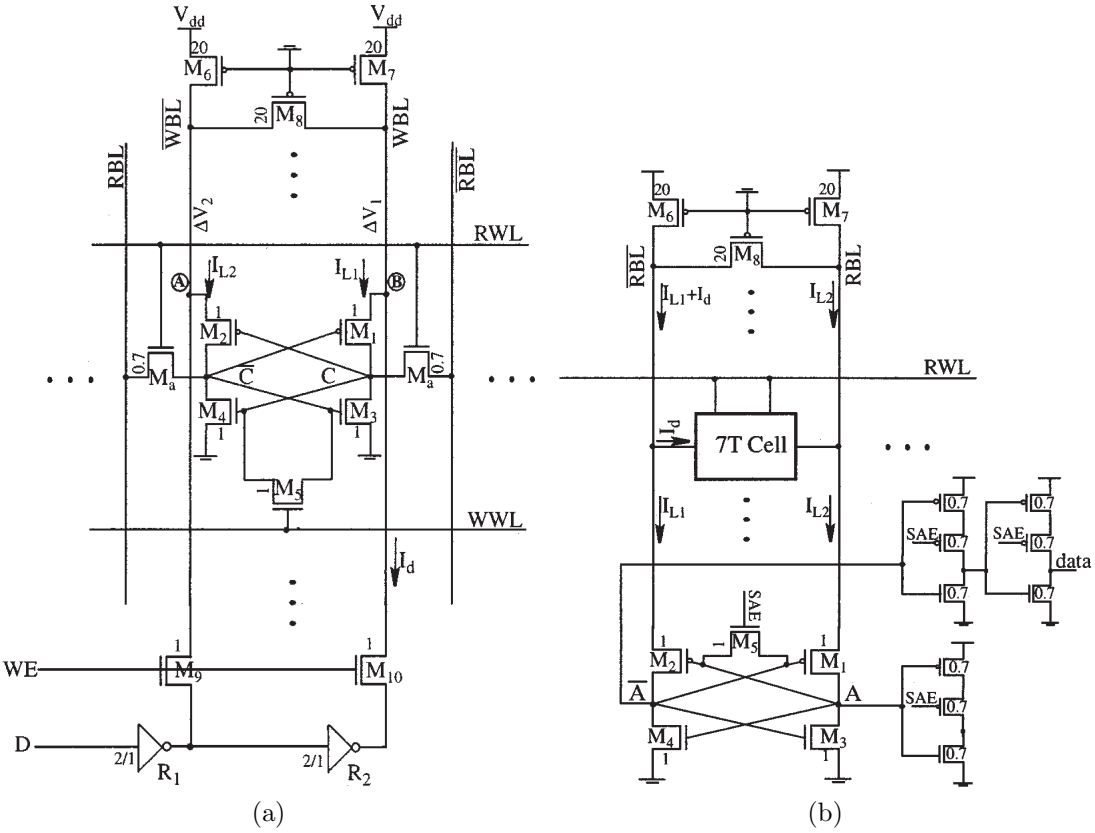


Figure 8: (a) SRAM with current-mode write. (b) SRAM with current-mode read.

Figure 8(a) is an SRAM with current-mode write and voltage-mode read. A current-mode write circuit generally consists of two back-to-back inverters as the voltage-mode SRAM does, a write driver circuit and a bitline load cicuit. The function of the bitline load circuit is to clamp the bitline voltage close to V_{DD}, avoiding voltage drop on it. One column of the memory array only has one load circuit and one write driver circuit. A more complicated write driver circuit is proposed in [29]. It contains a current conveyor that can create equivalent voltage on the two terminals of the cell so that current difference on the bitline can be injected into the back-to-back inverters. The basic idea of the current-mode SRAM is to initially bias the back-to-back CMOS inverters of the memory cell in the transient region to increase their voltage gain. Then the write driver creates a small voltage difference on the bitlines. The voltage difference causes different currents to be injected into the cell, which will be amplified to the full CMOS voltage due to the back-to-back inverters in their transient region. The voltage difference can be as low as a few milivolts when the supply voltage is 1.2V or 0.9V. For the same load capacitance, the charging or discharging time of the current-mode SRAM can be less than 1/10 of that of the traditional voltage-mode SRAM. Therefore, current-mode SRAM will certainly decrease the write dynamic power consumption of the memory at the cost of larger area in its write driver and bitline load circuit. There are also methods for current sensing in read circuits as shown in Figure 8(b). The sensing scheme is similar to that in write circuit, which is to convert the voltage difference of the internal node into different current on the bitline. The sense amplifier, which contains two back-to-back inverters, injects different currents and the internal two nodes are amplified to full CMOS voltage.

3.2 THE WORKING PRINCIPLE AND THE CHALLENGES OF CURRENT-MODE SRAM

Figure 9 shows the write circuit of the original current-mode SRAM we use in our design, i.e., connecting the bottom right red box to the cell [28], since the challenges and the innovations are both on the write circuit. It is the fundamental part of the multi-port current-mode

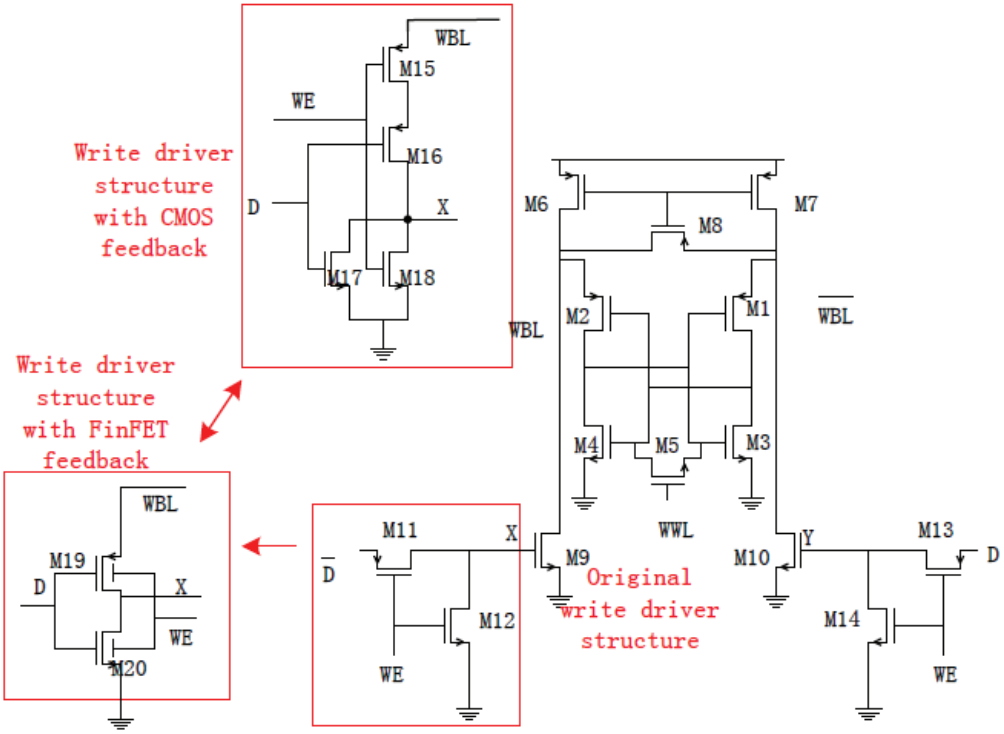


Figure 9: The proposed write circuit

SRAM [4]. The write operation can be performed within two phases: equalization and evaluation. In equalization, two storage nodes are equalized to an intermediate voltage and a current difference is generated on the two bitlines by switching either M9 or M10 on. Then, in evaluation, one of the storage nodes is set to '0' while the other is set to '1' by using the current difference on bitlines. One of the issues in this design is that the voltage of the bitline WBL is pulled down to a certain point when W9 is on, and other cells in the same column may be unstable if this voltage drop is greater than the threshold voltage of M2 in those cells, due to the sharing bitline.

3.3 OUR NOVEL FINFET CURRENT-MODE MULTIPORT SRAM

One novel method we propose aiming to address the challenge is shown in Figure 9. The top red box is the feedback circuit in planar CMOS, which is similar to a NAND gate.

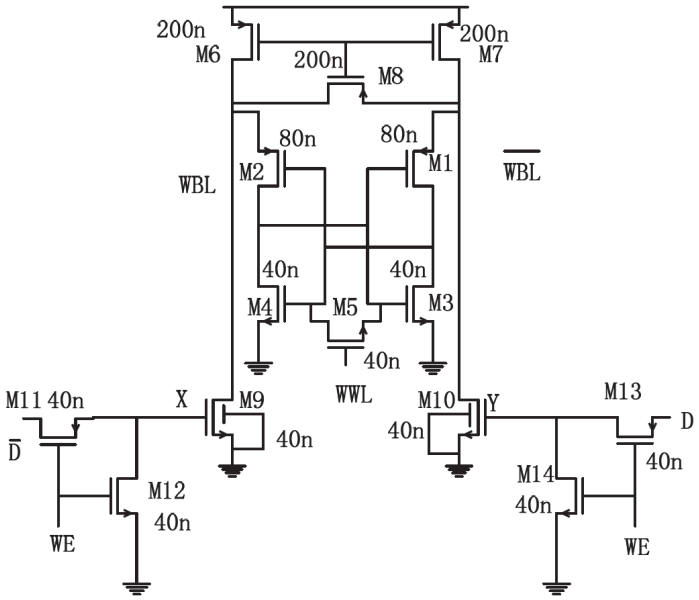
The difference between them is that its supply voltage is provided by the bitline. The idea behind this design is to use the bitline voltage to control the gate voltage of the write access transistor, M9. If voltage drops at the bitline, the gate voltage of M9 drops and results in a large resistance that prevents the bitline voltage from further dropping. To reduce the number of transistors, we use the transistor merging technique, which is a unique feature of FinFET. M17 and M18 can be replaced by a low-V_t transistor (M19) with back gate control; M15 and M16 can be replaced by another high-V_t transistor (M20) [30]. That can be done because high-Vt transistor will have low resistance if and only if both of the gates are activated and low-V_t transistor will have low resistance if either of the gates is activated.Those are similar to an AND function and an OR function, and therefore, a low-V_t FinFET transistor can replace two transistors in parallel and a high-V_t transistor can replace two transistors in series. Thus, this FinFET design not only adds no extra transistors, but also avoids using large transistors for the PMOS load transistors (M6, M7, and M8). [30] proposes a method to build high-Vt and low-Vt FinFET transistors. The threshold voltage of a FinFET threshold voltage is approximated by

$$V_t = -\phi_{ms} + \frac{Q_D}{C_{ox}} + V_{inv} + V^{QM} - V^{SCE} \tag{3.1}$$

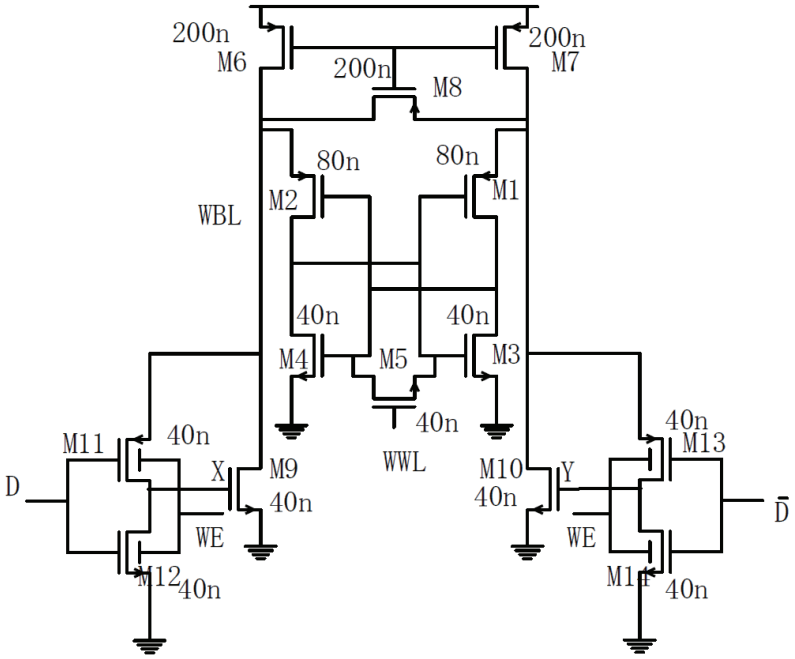
V^{SCE} models the short channel effect, and ϕ_{ms} is the potential difference between electrode and silicon. A high threshold voltage can be achieved only by manipulating the ϕms and V^{SCE} terms. V^{SCE} is mainly governed by the thickness of silicon. Therefore, decreasing work function and t_{si} can increase the threshold voltage. Our simulation is based on the range of those two parameters given in this paper and hence it is convincible.

Another method that we propose to avoid voltage drop on the bitline is to directly increase the on resistance of the bitline. This method is impossible in CMOS because the size of the transistor is in its minimum. However, in FinFET, we can use back-gate control to satisfy this requirement by decreasing the back gate voltage to increase the threshold voltage of the transistor. Figure 10(a) shows the method.

In fact, this method often performs better than the previous method because the bitline voltage of the former can be higher than that of the latter when all of the transistors are of normal size as shown in Figure 10 and the parameters of FinFETs in IG mode including



(a)



(b)

Figure 10: (a) The proposed current-mode SRAM with feedback in the write driver circuit. (b) The proposed current-mode SRAM with IG mode access transistors.

M9, M10 in Figure 10(a) and M11, M12, M13, M14 in Figure 10(b) are in the permitted range [30]. That is because in Figure 10(a), when the back gate of M9, M10 is tied to ground, the on resistance of that transistor is much bigger than a SG mode transistor, thus occupying more voltage drop. However, that case causes the decrease of bitline current, which reduces the speed of discharging one of the bitlines to create the difference of bitline voltage. However, in simulations we found that the magnitude of the voltage difference does not affect the speed of settling down of the voltage of the internal two nodes in evaluation stage when the lower one of the two bitline voltages is less than 1.1V for a supply voltage of 1.2V. If the evaluation stage starts when one of the bitline voltages discharges to above 1.1V, method 1 is a better choice.

Furthermore, in simulation of both the structures in Figure 10(a) and 10(b), the internal node that should settle down to 1.2V cannot reach this voltage unless the write enable signal is disabled because the voltage of the bitline on this side is somewhat less than 1.2V, and the transistor M8 in both structure equalizes the bitline voltage. That stops the use of the voltage of the internal node. Therefore, we eliminate M8 to de-equalize the bitline voltage. Although that will decrease one of the bitline voltages, we can use the two methods described in Figure 10 to compensate in practice.

For multi-port current-mode SRAM design, we choose the second structure as an example. Figure 11 shows the two-port current-mode SRAM. M9 and M10 are two IG-mode transistors, which combine two access transistors together in order to save area. The more ports there are, the more area can be saved. The performance of this two-port current-mode SRAM is like the structure in Figure 10(b) because when one port is disabled, the corresponding gate should be tied to ground. In [4], there is another method that has been proposed for multi-port design. This design also can save area because three ports share one access transistor. However, the number of ports tied to one transistor cannot be too large because the bottom enabled transistor should discharge all the nMOS transistors above it to make the gate voltage of the access transistor return to ground level, causing long latency of write. For our viewpoint, we can combine this method and the method in Figure 11 to design multi-port current-mode SRAM. In addition, as Figure 12 shows, the IG mode transistor can be implemented in this fashion to save area.

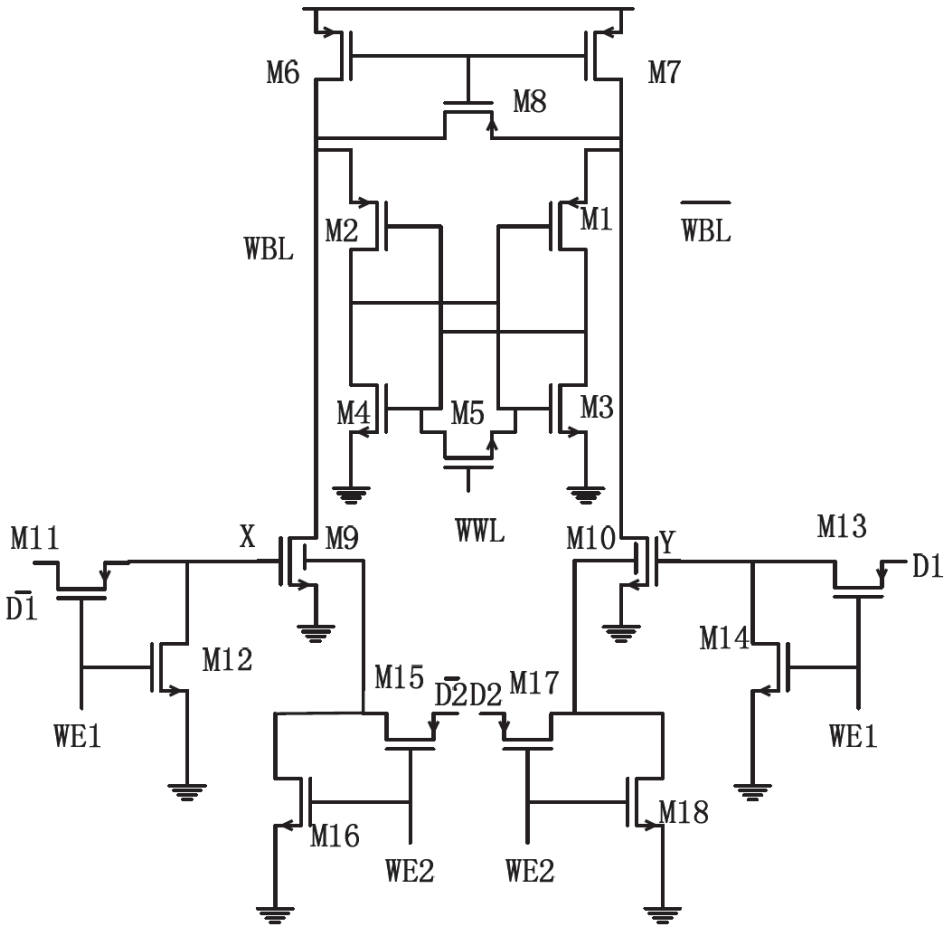


Figure 11: Proposed two-port current-mode SRAM.

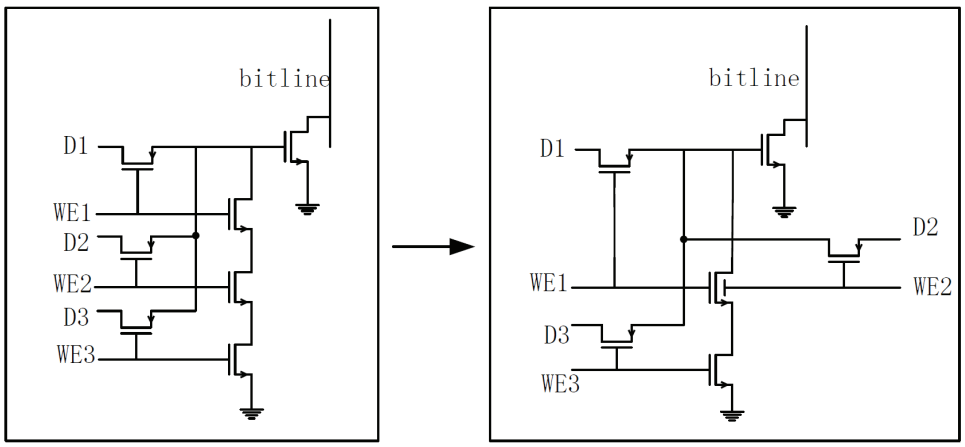


Figure 12: Using the merging transistors technique of FinFET to save area in the current-mode multi-port register file [4].

3.4 RESULT

Table 2 shows the result of the bitline that is pulled down by the corresponding open access transistor for the original circuit, method 1 and method 2. In simulation, transistor M8 is open to raise the voltage of the other bitline so that the internal node that should settle to '1' can quickly reach the final voltage after the write-enable signal is disabled. For example, for the condition of 40nm width for load transistors, using Method 1 with M8 closed, before when write-enable signal is enabled, the internal node with high voltage will be settled to 1.03V while without M8 closed, this node will be settled to 1.10V. Therefore, the time for the latter structure to settle down to 1.2V is much shorter than that of the former one. However, as table 2 shows, the voltage drop is bigger when M8 is open. That drop can be larger than the threshold voltage to cause the serious problem as mentioned when the bitline is below 0.89V with 1.2V as the supply voltage since the threshold voltage for UFDG transistors is 0.31V. Method 1 and Method 2 can prevent this phenomenon from happening when the width of load transistors of Method 1 is greater than 80nm and that of Method 2 is greater than 40nm. In Method 2, we use different thickness of silicon in the access transistors (T_{si}) to simulate since it directly affects the threshold voltage of those

Table 2: voltage drop on the bitline using original circuit, Method 1 and Method 2 when the width of load circuit is 40nm, 80nm and 200nm

		40nm	80nm	120nm
M8 is closed	Original structure	0.72V	0.92V	1.1V
	Method 1	0.92V	1.02V	1.12V
M8 is open	Original structure	0.488V	0.85V	1.06V
	Method 1	0.857V	0.974V	1.108V
	Method 2(T_{si}=0.014μm)	0.963V	1.084V	1.129V
	Method 2(T_{si}=0.005μm)	1.1V	1.12V	1.14V

transistors. Results in the table shows even if we choose the thickest one, the performance of that is better than Method 1. We also simulate the corresponding multi-port current-mode SRAM with two ports using the methods in Figure 11 and 12 (left circuit) respectively. The time for their bitline voltages to settle down in equalization stage is almost the same as the time of Method 2 used in single-port current-mode SRAM design. When using the right circuit in Figure 12 to mix transistors, the settled bitline voltages will be higher than the previous two methods, but that does not affect the settling down of the internal nodes in the following evaluation stage. As a result, we must choose a better structure, suitable parameters and sizes of the transistors to both avoid the bitline voltage to decrease by more than the threshold voltage of the top transistor and maintain the speed of equalization and evaluation stage.

3.5 SUMMARY

This chapter gives two methods with FinFET to relieve the voltage drop problem on bitlines of current-mode SRAM. One method is to create a feedback loop to prevent the voltage from further dropping. The other is to tie the back gate voltage to ground to directly increase the resistance of the access transistors to control the voltage drop. The first one has the advantage of flexibility, and the second one mostly have less voltage drop than the first one but with less flexibility.

4.0 1R/1W TWO-PORT PCM CELL

In this section, we will talk about the one read and one write two-port PCM cell that we designed. This two-port PCM cell is a novel structure because its two access transistors are tied to the bitlines instead of the ground. The advantage of this cell is to reduce the read conflicts with write request at the bank level in the network memory.

4.1 INTRODUCTION OF PHASE-CHANGE MEMORY

Phase-change memory (PCM), an emerging non-volatile memory, simply consists of two components, an access transistor and the chalcogenide material as the storage unit (GST). Generally, both CMOS transistor and PN diode can be used as the access transistor. Diode is sometimes a better choice because of its high effective current flow [31]. We need high current in write operation, and to produce such high current, CMOS transistor is not always stable. The GST can be programmed into two states: crystalline and amorphous.These two states are characterized by remarkably different resistance levels, where the amorphous chalcogenide material has the high resistance, usually in the MΩ range, and the crystalline state chalcogenide material has the low resistance, usually in the kΩ range [32].

There are three primary operations integral to the use of PCM in a modern memory system: read, SET, and RESET. The read operation loads the data from the memory to the processor or the cache hierarchy. The SET operation writes the bit '1' to the memory cell, i.e., the SET operation changes the state of the chalcogenide material in the cell to amorphous. In contrast, the RESET operation writes the bit '0' to the memory cell by changing the state of the chalcogenide material in the cell to polycrystalline.

A PCM cell can be read by simply sensing the current flow. Due to the large gap between the two resistance levels of the chalcogenide material, the sensing current flows of these two states differ by 3 or more orders of magnitude. The latency of the read operation in PCM cells is typically tens of nanoseconds.

In the write operation, the programming circuit of PCM applies different heat-time profiles to switch cells from one state to another. To RESET a PCM cell, a strong programming current pulse of short duration is required. The temperature of the chalcogenide material is raised by this programming pulse. After the chalcogenide material reaches the melting point, typically higher than 600°C, the programming pulse is quickly terminated. Subsequently, the small region of melted material cools quickly, resulting in the chalcogenide material programmed into the amorphous state. Since the region of the melted chalcogenide material is smaller, the required duration of the RESET programming pulse is short, about tens of nanoseconds. Thus, the RESET latency is typically similar to the read latency [33].

In contrast, to SET a PCM device, a long programming current pulse, which is weaker than the RESET programming current, is applied to program the cell from the amorphous state to the polycrystalline state. In the SET operation, the temperature of chalcogenide material should be raised above its crystallization temperature but below the melting point for a sufficient amount of time. As the crystallization rate is a function of temperature, given the variability of PCM cells within an array, reliable crystallization of PCM cells requires a programming pulse of hundreds of nanoseconds in duration [33]. Therefore, the SET latency is much larger than both the RESET latency and the read latency.

Phase-change memory has its priorities. As a non-volatile memory, the data retention time can be several months to even several years according to the capacity, the memory architecture and the number of write launched for PCM. Furthermore, scaling is not a big problem in PCM because the phase change materials (GST) undergoes excellent intrinsic scaling properties. Ultra-thin (up to 3 nm thick) phase change materials have also been shown to exhibit excellent data retention and cycling characteristics.

4.2 PCM METHODOLOGY AND ITS CHALLENGES FOR NETWORK PROCESSING

Modern network devices such as Internet routers have become highly dependent on scalable memory architectures. A large amount of data needs to be moved and managed in such devices, requiring significant memory capacity and bandwidth that increases with the line rate [34]. Therefore, memory systems in network devices must be capable of supporting fast read and write accesses at line rates, while also offering a large memory space necessary to maintain large data structures. DRAM has played a major role in supporting the demands on memory capacity and performance for network processing, largely in the form of hybrid SRAM/DRAM packet buffer [35, 36] and virtually pipelined memory architectures [37, 34]. However, scaling DRAM below 22nm is currently unknown [38], which makes DRAM less suitable for network processing in the "big data" era.

PCM, which has shown its scaling advantage, offers read latency close to that of DRAM and is a promising candidate to fill this scalability gap. Unfortunately, PCM is an asymmetrical read-write technology with a write latency is much longer than that of DRAM. The long write latency, usually 5–10× the read latency [39, 40], significantly increases bank conflicts over DRAM. To make things worse, read requests are latency critical in networking applications and cannot be scheduled with buffers like write requests. When PCM is used to implement virtually pipelined network memory, the long write latency of PCM requires longer fixed pipeline delay for both reads and writes. Simply put, the fixed pipeline delay is a linear function of PCM write latency (at least 10× equivalent DRAM pipelined memory). Thus, the asymmetrical write/read latency inherent to PCM remains the biggest challenge that has to be overcome in order to realize scalable PCM network memory.

4.3 TWO-PORT PCM CELL: MOTIVATION AND DESIGN

To solve the problem of asymmetrical write/read latency inherent to PCM for the network processing applications, we separate the write and read port. The two-port PCM cell

significantly reduces the probability of blocking at the bank and architecture levels and accelerate the read and write operation at the cell level because it bears the ability of serving simultaneous read and write at the cost of total area overhead. we will present necessary methods, such as voltage pumping and transistor size selection, to maintain the programming/sensing current requirement in dual- porting. We will also estimate the area overhead of the proposed cell design, and discuss the tradeoffs between area overhead and voltage pumping.

Our basic two-port PCM memory cell, which is illustrated in Fig. 13(b), consists of two access transistors and a GST storage material. Two bitlines and two wordlines are connected to two access transistors to compose a two-port (1R1W) design. One of these two ports supports only reads, while the other supports only writes. The two transistors are located at the crosspoints of bitlines and wordlines. Note that the figure illustrates our high performance cell with pMOS access transistors in the read and write ports; the use of nMOS access transistors as well tradeoffs to reduce cell area and lower power are discussed later in this section.

We have used SPICE simulations to validate the two-port PCM cell with the Predictive Technology Model (PTM) model for the access transistors [41]. In order to evaluate the PCM cell, we model the I-V curve of the GST in Verilog-A with the data from [39, 42]. The I-V curve is implemented in a lookup table approach. We set the Ovonic threshold switching (OTS) point as $I_{OTS} = 10\mu A, V_{OTS} = 1.14V$. We use a quadratic function to represent the curve when $I < I_{OTS}$, and a linear function when $I > I_{OTS}$, as illustrated in Fig. 13(c). The quadratic function represents the amorphous state and the linear function represents the crystalline state. The OTS point means if the current rises up to this point, the GST will change its state from amorphous to crystalline.

Since we place access transistors on the top of the GST, we expect that the voltage of bitlines needs to be increased in order to get the equivalent current when the cell is accessed. As voltage pumping for write access is common in PCM [39, 43], increasing the voltage in the write port is a practical approach for the two-port PCM cell. We summarize our results in Table 3, which indicates that the two-port PCM cell can achieve equivalent write performance to a conventional single-port PCM cell — 700μA set current and 1000μA reset

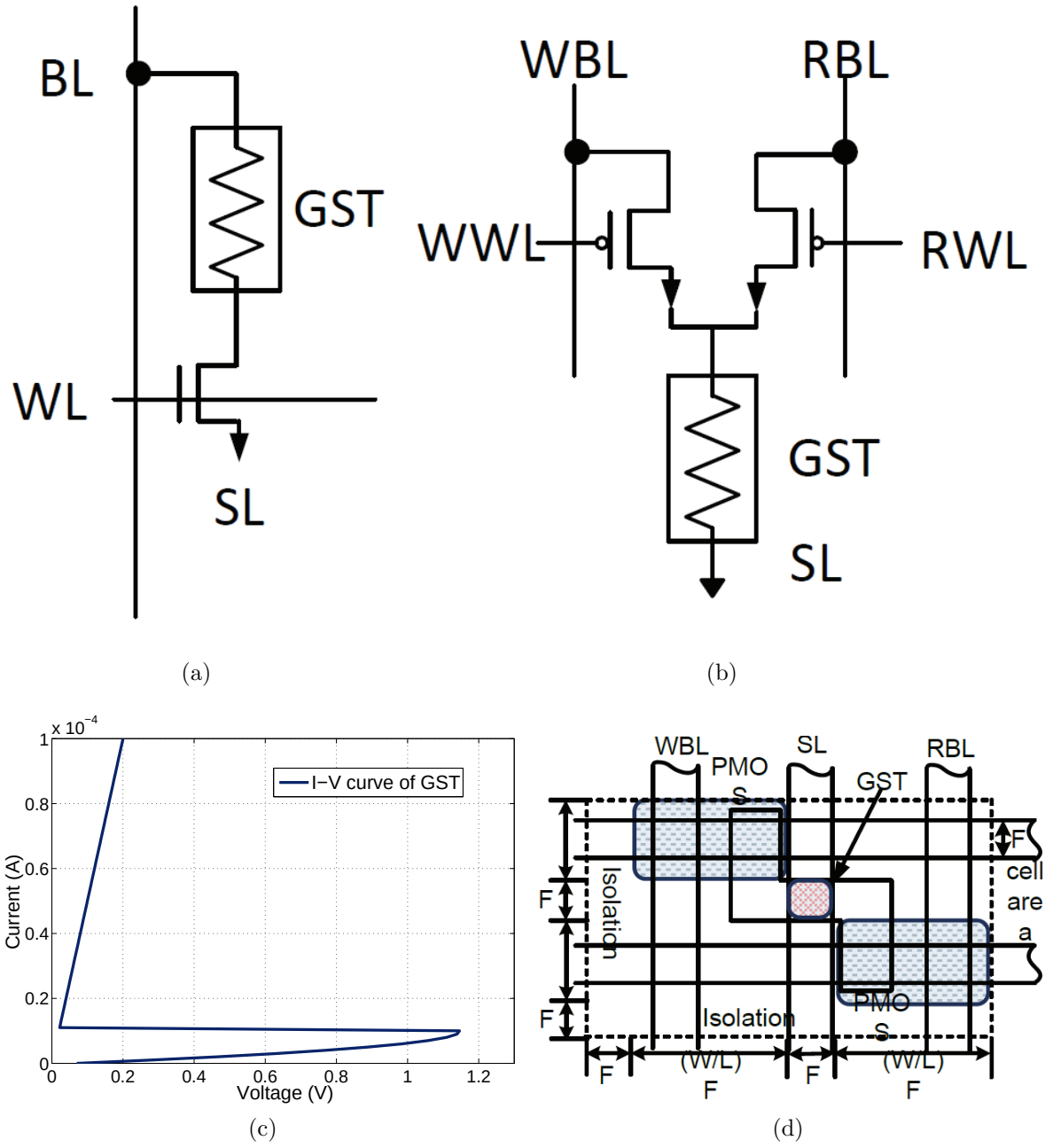


Figure 13: (a) The conventional single-port PCM cell, (b) the proposed two-port PCM cell, and (c) I-V curve of the GST model (d) PCM cell layout

Table 3: Voltage pumping for the write port

		90nm	65nm	45nm	32nm
V_{WBL}	Classical	4.06V	4.41V	4.75V	4.94V
	pMOS	5.63V	5.52V	5.41V	5.16V
	nMOS	4.39V	4.39V	4.39V	4.4V
V_{WWL}	Classical	2.56V	2.91V	3.25V	3.44V
	pMOS	1.5V	1.5V	1.5V	1.5V
	nMOS	2.87V	2.87V	2.87V	2.9V

current [39] — by boosting V_{WBL} in the write port from 4.06V to 5.63V in 90nm technology. Note that we use the 90nm PTM model as the reference to enable a fair comparison, since it is the closest model in PTM to the 100nm technology in [39]. We also set the W/L ratio of both pMOS and nMOS in our two-port cell to 4, while the nMOS in the classical PCM has the W/L ratio of 5. The advantage of W/L=4 for nMOS/pMOS is reduced cell size ($60F^2$ versus $72F^2$). The tradeoff is an increase in VDD (by 3.8%) to provide sufficient programming/sensing current using voltage pumping. We will discuss the tradeoff between the size of access transistors and the required voltage pumping in details later. Higher V_{WBL} is needed in pMOS to provide required set/reset programming current than that of nMOS, which can be achieved by voltage pumping. We also observe that the V_{WBL} necessary for the pMOS write access transistor decreases as the technology scales down, while the V_{WBL} for nMOS write access transistor increases. Moreover, in the write operation, both V_{WBL} and V_{WWL} of the nMOS access transistor need to pump to a certain level. Note however that after the write operation, the punch-through effect may occur if V_{WWL} drops before V_{WBL}. Thus, when using nMOS write access transistors, a voltage pumping control circuit is necessary to avoid the punch-through phenomenon. For this reason, we believe that using pMOS write access transistors is more practical than using nMOS write access transistors.

Meanwhile, in the PCM read operation, the read current should simultaneously be large enough to enable detection and small enough to avoid disturbance. Thus, in the conventional

single-port PCM cell, the V_{RBL} of bitline is set to 0.6V in the read operation. We investigate the extra voltage needed for the read port of our two-port PCM cell to obtain equivalent performance to [39], which is 5μA read current in amorphous state. We also summarize the results in Table 4, showing that the necessary V_{RBL} is 0.92V, compared to 0.9V in conventional cell, and pMOS transistor needs lower V_{RBL} and V_{RWL} than that of nMOS to ensure required read current. Since pMOS requires lower V_{RBL} than nMOS in read port, we select pMOS as the access transistor for the read port. It is worth mentioning that since the required V_{WBL} of nMOS in the write port is significantly lower than that of pMOS, nMOS can be used to design a low power two-port PCM cell.

Table 4: Voltage pumping for the read port

		90nm	65nm	45nm	32nm
V_{RBL}	Classical	0.9V	0.9V	0.9V	0.91V
	pMOS	0.92V	0.93V	0.95V	0.97V
	nMOS	1.5V	1.51V	1.52V	1.52V
V_{RWL}	Classical	0.6V	0.6V	0.6V	0.61V
	pMOS	0.3V	0.3V	0.3V	0.3V
	nMOS	1.2V	1.21V	1.22V	1.22V

Finally, we estimate the cell size of our two-port PCM cell by following the cell area model [44, 45]. The actual size of the pMOS is 2F $\times(W/L)$F. Including the isolation area, the memory cell size in the two-port PCM cell configuration is $6 \times 2(W/L + 1) = 60\text{F}^2$ ($0.486\mu\text{m}^2$) in 90nm technology, shown in Fig. 13(d). For a fair comparison, we estimate the cell size of the design in [39] with the cell area model in [44]. The estimated cell size is 18F^2 ($0.18\mu\text{m}^2$). Thus, the area overhead of our proposed two-port PCM cell is 1.7×, compared the single-port PCM cell in [39]. We also compare the tradeoff between voltage pumping and the size of pMOS access transistors, as illustrated in Table 5. When access transistors have the W/L ratio of 3, which means the cell size is $48F^2$, the required VDD is 5.97V/0.93V for the write/read port, compared to 5.63V/0.92V with access transistor of W/L = 4. Thus, if scalability is more important than power consumption, smaller access transistors should be selected; otherwise, larger access transistors can reduce power consumption.

Table 5: Voltage pumping and access transistors size

	W/L	90nm	65nm	45nm	32nm
Write	4	5.63	5.52V	5.41V	5.16V
V_{DD}	3	5.97V	5.79V	5.55V	5.33V
Read	4	0.92V	0.93V	0.95V	0.97V
V_{DD}	3	0.93V	0.94V	0.97V	0.99V

4.4 TWO-PORT PCM CELL FOR NETWORK MEMORY

After talking about the detailed design of the two-port PCM cell, we finally go to the architecture level and describe how the cell reduces the read delay of the memory banks [46].

Based on our proposed two-port PCM cell, we further two-port the PCM bank, as illustrated in Figure 14(c), which organizes two-port PCM cell arrays in blocks. The separate read/write port of the bank can significantly reduce the delay of a read/write request, due to reduction in the number of bank conflicts.The two-port PCM cell arrays, as shown in Figure 14(b), are organized in banks such that a block consist of four cell arrays. The write current driver circuits with charge pumps are connected only to the write port bitlines, while the read sense amplifier circuits only serve the read ports Figure 14(c). At the bank level, we use a write buffer to queue write accesses. When a read access is issued to a PCM bank, the memory controller of this bank first checks the write buffer to see if there is a pending write to the same row in the write buffer, or if a write has been issued to the write port. In either case, data forwarding is implemented and the read access is serviced without accessing the cell array. Thus, as long as the write buffer does not overflow, write requests can be buffered and retired without blocking any read request. Meanwhile, when a write request is blocked by an ongoing read access to the same page, the write request remains in the write buffer until the read is completed. Since the read latency of PCM is $5-10\times$ smaller than the write latency of PCM, this blocking is insignificant to the performance of network memory.

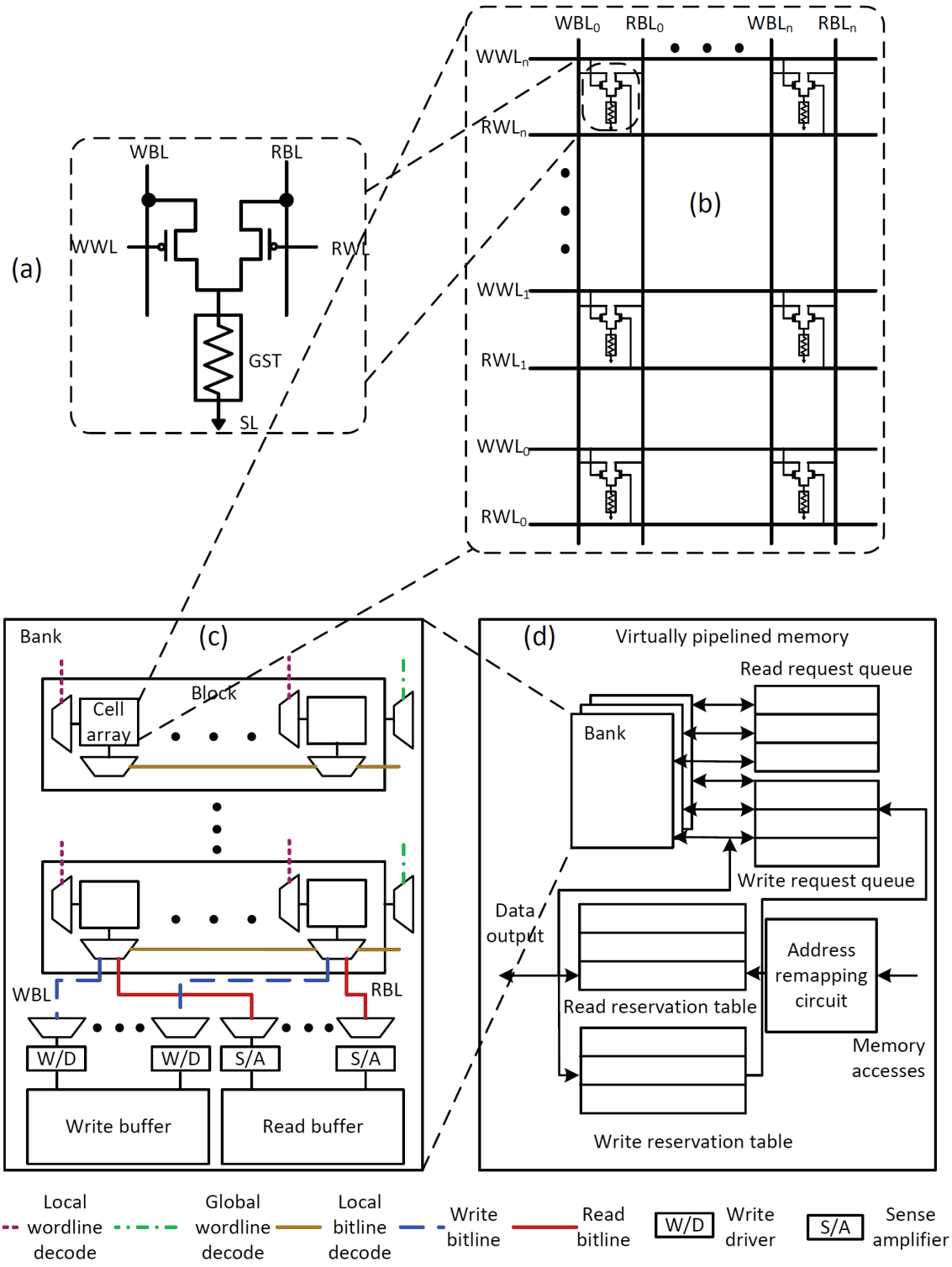


Figure 14: Proposed two-port PCM-based network memory: (a) cell schematic, (b) cell array, (c) two-port PCM bank, and (d) virtually pipelined memory architecture.

At the highest level, the virtually pipelined PCM memory is implemented as illustrated in Figure 14(d). It consists of a random address remapping function, a pair of reservation tables, the read/write request buffers, and the read/write tracking lookup table. The random address remapping function is realized to increase the bank parallelism. The dual-reservation table is to ensure there are no read request penalized. The write reservation table has $5-10\times$ entries that of the read one.We use buffers to queue the requests to each PCM bank. A write request buffer and a read request buffer, which are realized by SRAM, are associated to the corresponding ports of each PCM bank, to provide the fast network processing throughput. We also implement the write and the read tracking lookup table by using content addressable memory (CAM) for tracking the latest data update of a given memory address.

Our simulation framework considers three different applications: IP Security Protocol (IPSec), Flow Classification (flow class), and a IPV4 packet forwarding applications (IPV4-radix) from PacketBench [47]. These three applications represent various network processing applications: IPSec reads and modifies the packet payload, flow class is a classic network monitoring application, and IPV4-radix represents the most common applications in network processing: packet forwarding.

We simulate the network processor on the SimpleScalar simulator configured for an ARM core at 667MHz [48] and a 256-bank memory. The write/read latency of DRAM and the read latency of PCM are set to 40ns, and the write latency of PCM is set to 200ns. We use the traces collected from CAIDA's Equinix-Chicago monitor in 2003, 2008, and 2011. In our simulation, we compare the performance of three pipelined memory architectures: single-port DRAM, single-port PCM and our proposed two-port PCM, by evaluating four different metrics: memory access rate, port utilization, average number of waiting request, and the average delay.

Simulation results show that this memory architecture can reduce the expected read (write) delay by 12–40$\times$ (up to 14%) over conventional single-port PCM for 1.1–1.7$\times$ overhead. And the sum of the number of waiting requests in the write/read ports is only 33.1% to 49.8% as many as that of the single port PCM.

4.5 SUMMARY

We proposed and comprehensively evaluated the two-port PCM cell design in terms of programming current, necessary voltage pumping for access transistors, and area overhead. We conclude that both pMOS and nMOS on top is suitable for two-port design in the normal working situation of the access transistor. pMOS on top is better than nMOS on top due to its less pumping voltage when other parameters are the same. Analysis done in architecture level shows that this 1R1W PCM substrate can significantly reduce the expected delay of a read access for networking applications. Furthermore, it also reduces the number of waiting requests at the bank level, leading to a smaller buffer size.

5.0 CONCLUSIONS AND FUTURE WORK

5.1 CONCLUSIONS

In recent years, scaling down of transistors and the higher requirements on the performance such as power, area and speed pose great challenges on the design of memories. In this thesis, we describe and evaluate novel memory designs for multi-port on-chip and off-chip use in advanced computer architectures. Multi-porting is essential for caches and shared-data systems. It can significantly increase the memory access throughput.

For on-chip memories, several FinFET multi-port SRAMs are proposed. Our evaluations of read and write acceleration in those different structures illustrate the impact on read/write performance, leakage current, and cell stability. Based on simulation results with the PTM FinFET model, single-ended multi-port FinFET SRAM with isolated read ports is a good choice for multi-port design, since for similar leakage current, write time, and 9% area overhead, it performs better in read operation, offers higher flexibility in the configuration of read acceleration, and provides better cell stability than double-ended multi-port FinFET SRAM. Compared with corresponding CMOS SRAMs, FinFET SRAMs displays a better performance in stability and standby power due to its advantage of suppressing short channel effect.

Besides using FinFET in voltage-mode multi-port SRAM design, we also propose two novel structures in current-mode multi-port SRAM and IG mode FinFET is also applied by merging parallel transistors to save area and improve the performance. The problem of the voltage drop on the bitline is substantially improved. When the width of load transistors is 40nm, Method 1 produces 41.67% reduce on the voltage drop with transistor M8 on and 51.8% reduce with M8 off. Method 2 usually even performs better than Method 1.

Furthermore, multi-porting by merging IG mode transistors to reduce area does not affect much on the speed of write access.

For off-chip memories, a two-port non-volatile PCM is proposed. Instead of the traditional structure with the access nMOS transistor at the bottom, we put the access transistor on the top and compare the performance of both nMOS and pMOS transistor as the access transistor. We comprehensively evaluated the two-port cell design in terms of programming current, necessary voltage pumping for access transistors, and area overhead. We come to the conclusion that the pMOS access transistor on the top is more favorable because it requires less supply voltage and write driver voltage in most technologies we use. Compared with the single-port cell, the two-port cell only has an 1.7× increase on the area overhead. Analysis done in architecture level shows that this 1R1W PCM substrate can significantly reduce the delay of a read access for networking applications. Furthermore, it also reduces the number of waiting requests at the bank level, leading to a smaller buffer size.

5.2 FUTURE WORK

Spin-transfer torque random access memory (STT-RAM) is another off-chip non-volatile memory as PCM. Although extensive research has been performed on this memory, multi-port STT-RAM has not been proposed yet. Multi-porting the STT-RAM can relieve its shortcoming of long write access latency overhead. Correct modeling MTJ (the storage component) is challenging because the state of the MTJ will change if the density of current reaches the critical values. PCM and STT-RAM has similar cell structures. Therefore, we also can design multi-port STT-RAM using the design approach similar to that described in this thesis.

Another direction for further research is the use of FinFETs to improve the design of multi-port Schmitt-Trigger SRAM. Schmitt-Trigger SRAM has an outstanding advantage over traditional 6T SRAM by its cell stability, but the cost is its large area. The key idea is to reduce the area by mixing two SG FinFET transistors into one IG mode transistor, and by extending this to every port.

BIBLIOGRAPHY

[1] N. Tzartzanis, "Static memory design," in *High-Performance Energy-Efficient Microprocessor Design* (V. G. Oklobdzija and R. K. Krishnamurthy, eds.), ch. 4, Springer, 2002.

[2] J. Singh, D. S. Aswar, S. P. Mohanty, and D. K. Pradhan, "A 2-port 6T SRAM bitcell design with multi-port capabilities at reduced area overhead," in *Proc. Intl. Symposium on Quality Electronic Design*, pp. 131–138, 2010.

[3] N.-H. E. Weste and D. Harris, "Array subsystems," in *CMOS VLSI design: a circuits and systems perspective (4th edition)* (M. Goldstein and M. Suarez-Rivas, eds.), ch. 11, Pearson and Addison-Wesley, 2012.

[4] S. Khelleh, Z. Li, and F. Wang, "Design of a high-speed low-power multiport register file," in *Microelectronics and Electronics. PrimeAsia*, pp. 408–411, 2009.

[5] M. K. Qureshi, s. Gurumurthi, and B. Rajendran, *Phase Change Memory: from devices to systems*. Morgan and Claypool Publishers, 2011.

[6] F. D'Agostino and D. Quercia, "Short-channel effects in MOSFET," project report, Department of ECE, University of Cambridge, "December" 2000.

[7] "Process variation of semiconductor (Wikipedia)." Please visit the URL http://http://en.wikipedia.org/wiki/Process_variation_(semiconductor) for further details.

[8] X. Yang and K. Mohanram, "Robust 6T Si tunneling transistor SRAM design," in *Design, Automation and Test in Europe Conference and Exhibition*, pp. 1–6, 2011.

[9] S. Chen and C. Wang, "Single-ended disturb-free 5T loadless SRAM cell using 90 nm CMOS process," in *IEEE International Conference on IC Design and Technology*, pp. 1–4, 2012.

[10] Y. Tseng, Y. Zhang, L. Okamura, and T. Yoshihara, "A new 7-transistor SRAM cell design with high read stability," in *Intl Conf on Electronic Devices, Systems and Applications*, pp. 43–47, 2010.

[11] C. Hsieh, M. Fan, V. Hu, P. Su, and C. Chuang, "Independently-controlled-gate FinFET schmitt trigger sub-threshold SRAMs," in *IEEE International SOI Conference*, pp. 1–2, 2010.

[12] Z. Guo, S. Balasubramanian, R. Zlatanovici, T. King, and B. Nikolic, "FinFET-based SRAM design," in *Proc. Intl. Symposium on Low Power Electronics and Design*, pp. 2–7, 2005.

[13] P. Manoilov and P. Krivoshieva, "Shared memory design for multicore systems," in *International Scientific Conference Computer Science*, pp. 302–307, 2008.

[14] M. Golden and H. Partovi, "A 500 mhz, write-bypassed, 88-entry, 90-bit register file," in *Proc. Symposium on VLSI circuits*, pp. 105–108, 1999.

[15] E. S. Fetzer and J. T. Orton, "A fully-bypassed 6-issue integer datapath and register file on an Itanium microprocessor," in *Proc. Intl. Solid-state Circuits Conference*, pp. 421–478, 2002.

[16] X. Dong, N. P. Jouppi, and Y. Xie, "PCRAMsim: System-level performance, energy, and area modeling for phase-change RAM," in *Proc. Intl. Conference Computer-aided Design*, pp. 269–275, 2009.

[17] K. Kim, C.-T. Chuang, J. Kuang, H. Ngo, and K. Nowka, "Low-power high-performance asymmetrical double-gate circuits using back-gate-controlled wide-tunable-range diode voltage," *IEEE Trans. Electron Devices*, vol. 54, pp. 2263–2268, September 2007.

[18] R. Joshi, K. Kim, and R. Kanj, "FinFET SRAM design," in *International Conference on VLSI Design*, pp. 440–445, 2010.

[19] Y. Chen, C. Hsieh, M. Fan, V. P. Hu, P. Su, and C. Chuang, "Disturb-free independently-controlled-gate 7T FinFET SRAM cell," in *Intl. Symposium on VLSI Technology, Systems and Applications*, pp. 1–2, 2011.

[20] N. Tzartzanis, W. W. Walker, H. Nguyen, and A. Inoue, "A 34word x 64b 10R/6W write-through self-timed dual-supply-voltage register file," in *Proc. Intl. Solid-state Circuits Conference*, pp. 338–357, 2002.

[21] H. Bajwa and X. Chen, *Area-Efficient Dual-Port Memory Architecture for Multi-Core Processors*. PhD thesis, City University of New York, 2007.

[22] T. Suzuki, H. Yamauchi, Y. Yamagami, K. Satomi, and H. Akamatsu, "A stable 2-port SRAM cell design against simultaneously read/write-disturbed accesses," *IEEE Journal of Solid-state Circuits*, vol. 43, pp. 2109–2119, September 2008.

[23] M. Yoshimoto, K. Anami, H. Shinohara, T. Yoshihara, H. Takagi, S.Nagao, S. Kayano, and T. Nakano, "A divided word-line structure in the static RAM and its application to a 64 K full CMOS RAM," *IEEE Journal of Solid-state Circuits*, vol. 18, pp. 479–485, October 1983.

[24] SOI Group of University of Florida, *UFDG MOSFET Model User Guide (Linux Version 3.71)*, 2003. Please visit the URL http://www.soi.tec.ufl.edu for further details.

[25] "PTM-MG multi-gate model," 2012. Please visit the URL http://ptm.asu.edu/ for further details.

[26] S. Yaldiz, U. Arslan, X. Li, and L. Pileggi, "Efficient statistical analysis of read timing failures in SRAM circuits," in *Proc. Intl. Symposium on Quality Electronic Design*, pp. 617–621, 2009.

[27] J. Wang, S. Nalam, and B. H. Calhoun, "Analyzing static and dynamic write margin for nanometer SRAMs," in *Proc. Intl. Symposium on Low Power Electronics and Design*, pp. 129–134, 2008.

[28] M. M. Khelleh and M. I. Elmasry, "Circuit techniques for high-speed and low-power multi-port SRAMs," in *Intl. Conference on ASIC*, pp. 157–161, 1998.

[29] J. Wang, P. Yang, and W. Tseng, "Low-power embedded SRAM macros with current-mode read/write operations," in *Proc. Intl. Symposium on Low Power Electronics and Design*, pp. 282–287, 1998.

[30] M. Rostami and K. Mohanram, "Dual-vth independent-gate FinFETs for low power logic circuits," *IEEE Trans. Computer-aided Design*, vol. 30, pp. 337–349, March 2011.

[31] L. Li, K. Lui, K. C. Kwong, J. He, and M. Chan, "Comparison of PN diodes and FETs as phase change memory (PCM) driving devices," in *Intl. Conference on Solid-State and Integrated-Circuit Technology*, pp. 928–931, 2008.

[32] J. Yue and Y. Zhu, "Accelerating write by exploiting PCM asymmetries," in *Intl. Symposium on High-performance Computer Architecture*, 2013.

[33] M. K. Qureshi, M. Franceschini, A. Jagmonhan, and L. Lastras, "Preset: Improving performance of phase change memories by exploiting asymmetry in write time," in *Proc. Intl. Symposium on Computer Architecture*, 2012.

[34] B. Agrawal and T. Sherwood, "Virtually pipelined network memory," in *Proc. Intl. Symposium on Microarchitecture*, pp. 197–207, 2006.

[35] S. Iyer, R. R. Kompella, and N. McKeown, "Designing packet buffers for router line cards," tech. rep., Stanford University, 2002.

[36] J. Garcia, J. Corbal, L. Cerda, and M. Valero, "Design and implementation of high performance memory systems of future packet buffers," in *Proc. Intl. Symposium on Microarchitecture*, 2003.

[37] H. Wang *et al.*, "Design and analysis of a robust pipelined memory system," in *Proc. Intl. Conference on Computer Communications*, 2010.

[38] "International Technology Roadmap for Semiconductors," 2011.

[39] S. Kang *et al.*, "A 0.1-μm 1.8-V 256-Mb phase-change random access memory (PRAM) with 66-mhz sysnchronous burst-read operation," *IEEE Journal of Solid-state Circuits*, vol. 42, no. 1, pp. 210–218, 2007.

[40] I. Song *et al.*, "A 20nm 1.8V 8GB PRAM with 40MB/s program bandwidth," in *Proc. Intl. Solid-state Circuits Conference*, 2012.

[41] "Predictive Technology Model," 2012. Available at: http://ptm.asu.edu/.

[42] X. Li *et al.*, "An SPICE model for phase-change memory simulations," *Journal of Semiconductors*, vol. 32, no. 9, pp. 1–4, 2011.

[43] K.-J. Lee *et al.*, "A 90 nm 1.8 v 512 Mb diode-switch PRAM with 266 MB/s read throughput," *JSSC*, vol. 43, no. 1, pp. 150–162, 2008.

[44] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "A low-power high-performance current-mode multiport SRAM," *IEEE Trans. Computer-aided Design*, vol. 31, no. 7, pp. 994–1007, 2012.

[45] F. Fishburn, B. Busch, J. Dale, *et al.*, "A 78nm 6F^2 DRAM technology for multigigabit densities," in *Proc. Symposium on VLSI Technology Digest of Technical Papers*, pp. 28–29, 2004.

[46] J. Li, D. Dgien, N. A. Hunter, Y. Zhao, and K. Mohanram, "Dual-port PCM architecture for network processing," in *Non-Volatile Memories Workshop*, 2013.

[47] R. Ramaswamy and T. Wolf, "PacketBench: A tool for workload characterization of network processing," in *IEEE Annual Workshop on Workload Characterization*, pp. 42–50, 2003.

[48] "Intel IXP4XX product line of network processors," 2012. Available at: http://www.intel.com/p/en_US/embedded/hwsw/hardware/ixp-4xx.